The Validity of Using Holistic Scoring to Evaluate Writing: A Critical Overview

Davida Charney, Carnegie-Mellon University

Abstract. Teachers, administrators, testing agencies, and researchers all need a valid, reliable method of assessing writing ability. Each group has turned to holistic ratings of writing samples as a reliable qualitative procedure for responding to the essential features of writing. Yet the validity of holistic ratings has never been convincingly demonstrated. This paper analyzes the implicit requirements for achieving reliable results from holistic ratings and argues that these conditions bring the validity of the ratings into doubt. The research available suggests that even in carefully supervised rating sessions, holistic ratings may be unduly influenced by superficial features of the writing samples. Those who use holistic ratings to evaluate writing ability need to give more serious attention to the validity of the scores that result.

Holistic Rating as a Reliable and Theoretically Valid Procedure

Many people in the field of rhetoric and composition would agree with Hirsch (1977) that the assessment of writing ability is the "single most important snag to practical progress in composition teaching and research." Finding a method of measuring the writing ability of an individual or a group is not only of practical importance for carrying out administrative functions in schools, but is also necessary for carrying out research on composing. Teachers, administrators and researchers have employed a variety of assessment procedures over the years. However, they have continually had difficulty in finding a method of assessment that is both reliable and valid. As in other areas of research and testing, these are the two necessary criteria for any measurement. 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. A *valid* measurement assesses what it claims to assess. So, a valid writing assessment would be sensitive to a writer's "true" abilities. Time and again, the methods that have been employed to measure writing ability have been criticized as either unreliable or invalid.

The author would like to express her gratitude to Brenda Rosen, Patricia Sullivan, David Kaufer, the anonymous reviewers for *RTE*, and especially Richard Enos, whose comments materially aided the preparation of this manuscript.

Current methods of measuring writing ability can be characterized as either quantitative or qualitative. Consider the quantitative methods first. These methods, which have also been referred to as "objective" or "indirect" methods, are often used in conjunction with standardized tests, such as the verbal sections of the SAT's. Standardized tests normally assess students' ability to distinguish between standard and non-standard English, or their ability to choose the most "correct" or the most "mature" alternative to a defective construction. In a typical objective test question, students are provided with a faulty sentence and are asked to select the best correction. By varying the test items and the types of faults, the students' proficiency with a range of writing skills can be tested, including diction, spelling, grammar, punctuation, syntax, sentence order and some aspects of style. Quantitative methods can also be used in conjunction with essay tests, by counting such things as occurrences of grammatical errors, the number of t-units, or the number of uncommon vocabulary items in a writing sample. Again, the focus of the assessment is the students' maturity, as reflected in their mastery of writing conventions.

Although many quantitative methods are statistically reliable, teachers and administrators have almost universally rejected them as primary measures of writing ability on the grounds that they are invalid. The critics argue that, although the particular skills that quantitative methods focus on may be necessary in order to write well, proficiency in these skills does not indicate writing ability *per se*. McColly (1970), for example, states: ". . . . the tests simply are not measures of writing. For the purposes of judging writing ability, they therefore should be ignored." For similar views, see Lloyd-Jones (1977), Cooper (1977), Odell (1981), and Gere (1980).

To see whether this criticism is justified, consider the two ways in which the validity of a measure can be established. First, one can claim that a measure has "predictive validity." If the results of the measure correlate with the results of another measure then the two measures may be considered equally valid. Thus, in order for the proponents of a standardized test to appeal to its predictive validity, they must show that the test results correlate with a "criterion," a previously validated measure of writing ability. At one time, the most commonly used criterion was grades in college English courses. Since some quantitative test scores correlated with college grades, including grades in freshman writing courses, these tests were accorded predictive validity. For one recent study in which quantitative scores were correlated with grades, see Culpepper & Ramsdell (1982). The choice of the criterion is quite important for deciding how much weight to give a finding of predictive validity. Godshalk, Swineford and Coffman (1966) point out that grades in freshman English courses may not be a satisfactory criterion, because such grades may not be based entirely on writing skill. Testing agencies

have since attempted to find a more acceptable criterion, but a discussion of their success will be deferred.

The objections raised by McColly (1970) and others concern a second kind of validity: the "face validity," or the reasonableness of the method. Does the method supply an assessment that is based on consistent application of acceptable criteria? Odell (1981) argues that the criteria applied in quantitative methods are unacceptable for assessing writing ability because they are too limited. He proposes to define competence in writing as "the ability to discover what one wishes to say and to convey one's message through language, syntax, and content that are appropriate for one's audience and purpose" (Odell, 1981, p. 103). If writing competence is defined in this way, then quantitative methods fail to apply appropriate criteria for assessing it. They cannot test a student's ability to generate alternative constructions and to select those that are most appropriate to a particular purpose and audience. At best, quantitative methods test whether a student can identify constructions couched in some generic "plain style." They are insensitive to a student's ability to write cogent, coherent and fluent prose. It is possible that writers with greater writing competence will also display greater mastery of writing conventions. If so, then the quantitative methods may in fact separate students according to writing competence. However, the critics of quantitative methods are asking for a sorting of writers that does not depend on this connection. For them to accord a measure face validity, it must assess writing skills beyond mastery of conventions.

Many have argued that the *qualitative* evaluation of writing samples offers an intuitively more valid approach than quantitative methods. In a qualitative method, students generate text about some fixed topic, and readers evaluate the writing samples as purposive messages. Qualitative methods thus allow for assessment of high level writing skills and therefore seem to apply more valid criteria. However, for many years qualitative methods suffered from a lack of reliability. Study after study demonstrated that readers who evaluate writing samples apply widely varying standards. Under normal reading conditions, even experienced teachers of writing will disagree strongly over whether a given piece of writing is good or not, or which of two writing samples is better. (For reviews of this research, see Godshalk, et al., 1966 and McColly, 1970.) As a result of the unpredictable variations in the ratings, qualitative assessments of writing samples could not be trusted as accurate. Testers and researchers turned to holistic ratings in the 1950s and 1960s when it was found that raters who were trained in this method of reading and evaluating writing samples could produce reliable results.

Holistic rating is a quick, impressionistic qualitative procedure for sorting or ranking samples of writing. It is not designed to correct or edit a piece, or to diagnose its weaknesses. Instead, it is a set of procedures for assigning a value to a writing sample according to previously established criteria.

There are various procedures in use for arriving at a holistic rating. Holistic ratings may be assigned simply on the basis of the total impression a piece makes on a reader. This method is known as "General Impression Marking." Or the holistic rating may be based on an explicit scoring guide, a list of specific linguistic, rhetorical or informational features of writing that the reader keeps in mind while rating the piece. In keeping with Cooper's (1977) terminology, as long as the reader is not required to count the occurrences of surface text features, the evaluation can be considered holistic. Testing agencies, such as the Educational Testing Service (ETS), and numerous schools now routinely assess writing samples with holistic ratings. In addition, researchers of writing pedagogy frequently use holistic ratings as dependent measures to evaluate teaching methods (e.g., Sanders & Littlefield, 1975; Clifford, 1981; Hilgers, 1980; Moslemi, 1975; Davis, 1979).

Despite such widespread use, the question of whether holistic ratings produce accurate assessments of "true writing ability" has very often been begged; their validity is asserted but has never been convincingly demonstrated. Quantitative methods were scrutinized for predictive and face validity. Qualitative methods, including holistic scoring, should be treated the same way. This paper will argue that on both counts, the validity of holistic scoring remains an open question. Before beginning that discussion, however, there is one class of arguments against the face validity of holistic ratings which I would like to dismiss.

All of the testers and researchers who use holistic ratings implicitly accept the idea that writing ability can be inferred from an end-product of the writing process. It might be argued that the evaluation of writing ability should take the individual's writing *process* into account, and that product-based methods of assessment, such as holistic ratings, are therefore invalid *a priori*. This kind of argument assumes that if holistic ratings are invalid for one purpose, namely for assessing the student's writing process, then they are invalid for all purposes. Flower and Hayes (in press) take a more reasonable approach. They point out that the purpose of evaluating a student's writing process is diagnosis. Product-based evaluations, including both quantitative and qualitative methods, do not produce diagnoses (although Odell [1981] believes that some methods of holistic rating can be modified to do so). Instead they produce summative statistics which compare the abilities of individuals or of groups of writers. These statistics will not be useful to the teacher trying to identify a given student's writing problems, but they will be useful to administrators and researchers who wish to predict whether a given student will pass Freshman Composition or to decide whether a group of students has benefited from a particular writing class. Product-based evaluations are the only feasible methods available for testing and for research projects involving large numbers of writers. As long as the limitations of the results, in terms of validity and reliability, are understood and re-

spected, there is no reason to reject holistic evaluations, or other product-based evaluations, out of hand.

The Practical Validity and Reliability of Holistic Ratings: A Set of Conditions

There are several different kinds of holistic rating procedures in use, but the common assumption behind all of them is that a valid assessment of writing ability includes a natural human response to a writing sample. If readers can be trained to respond in a consistent, acceptable way, then the ratings will be reliable and valid. The requirements for achieving this condition are fairly complex. Those who use holistic scoring assume that the assessments will be valid and reliable:

> if the design of the training and rating sessions takes the factors necessary for reliability into account;
>
> if the readers are qualified, and come from similar backgrounds;
>
> if the readers are "calibrated," that is, trained to conform to agreed upon criteria of judgment;
>
> if the criteria, which either are supplied to the readers in the form of a rating guide, or are decided upon by the readers as a group, are appropriate; and
>
> if readers work quickly, usually under supervision.

At first sight, each condition seems reasonable, and, in isolation, each condition has been convincingly defended as leading to a higher level of statistical reliability. Yet, in practice, these conditions are not imposed separately; in any given rating session, they interact. We shall see that when the conditions are considered more closely, they add up to a paradoxical set of requirements. In particular, the reliability of the ratings is contingent upon conditions that, in concert, may vitiate the validity of the method.

Issues of Design: Choice of Topic and Rating Procedure

The first condition for validity and reliability concerns the effect of the design of the test or experiment on the statistical reliability of a set of holistic ratings. Statistical reliability can be calculated in two ways: reading reliability and score reliability. Reading reliability estimates the probability that another group of competent readers would produce a comparable set of ratings for the test papers. The more the actual group of readers agrees in its assessments, the higher the correlation. If the assessment of a writer's ability is based on ratings of more than one writing sample, then a score reliability

can be calculated. The score reliability figure estimates the probability that a second set of samples produced by the same student would be rated in a comparable way by a new set of readers. Both reading and score reliability procedures were used by Godshalk et al. (1966) for ETS.

A list of factors in the design of a test or study affects the statistical reliability of the ratings. The list includes: the number of separate readings of each writing sample, the number of writing samples evaluated per student, the writing topic, the size of the rating scale, the consistency with which the readers are trained, the conditions under which the papers are read, and so on (Nold and Freedman, 1977; Freedman, 1981; McColly, 1970; Godshalk et al., 1966; Stiggins, 1982). Many of these factors can be adequately controlled by designing the study carefully and monitoring the training and rating sessions. Some of these factors, however, crucially affect the credibility of the test or study as a valid test of writing ability. For example, one important issue is the selection of writing topics. Specifically, should writing samples representing different aims of discourse be compared?

The proper procedure for selecting topics is a matter of controversy. Testmakers disagree over whether topics should be wide open or narrowly defined. If topics are wide open, then each writer can write about an aspect of the topic that is familiar. One wide-open topic, used by Godshalk et al. (1966), asked students to write an imaginative story about an experience (as observer or participant) or about a commonplace, inanimate object. Another topic, used by Culpepper and Ramsdell (1982), asked students to "Select a character, from history or fiction, and discuss two or three qualities that made that person remarkable." Because the task is less structured, students writing on an open-ended topic may interpret it in very different ways. Some writers may set themselves tasks that are too easy, and others tasks that are too hard. Would it be fair to treat these essays alike? Nold (in press) argues that some discourse aims, such as persuasion and exposition, are harder to achieve than others, such as simple narration. She believes, therefore, that writing samples with different aims should not be rated equally. Furthermore, McColly (1970) mentions a widespread belief that a topic which allows great freedom of response lowers the validity and reliability of the ratings. If Nold and McColly are correct, then topics should be selected which carefully define the aim of the discourse to be produced, and only topics with the same aim should be compared. For example, Freedman (1979) employed topics which were designed to elicit argumentative discourse. One of her topics reads as follows: "President Ford gave Nixon an 'unconditional pardon.' Do you agree or disagree with Ford's decision? Give reasons for taking your position" (p. 329).

In addition to requiring that topics specify the aim of the discourse, McColly (1970) argues that topics should be chosen which hold the effects of knowledge constant. The assessment of a student's writing ability, especially

in an impromptu writing situation, should not depend on how familiar the writer is with the topic that is selected. Myers (1980) also discusses many of the problems related to topic selection, and recommends that topics be field-tested and reviewed for problems of focus, special knowledge, open-endedness and grade-level differences. To meet these recommendations, test questions can be limited to general knowledge, for example, by asking students to write about familiar proverbs, or can require students to make use of information provided in the exam. Lloyd-Jones (1977) reviews the disparate criteria that have been recommended for essay topics and correctly points out that there is a trade-off between attempting to elicit samples of particular kinds of discourse, and permitting each student to write on a topic which is interesting and familiar. "The more one restricts the situation in order to define a purpose and stimulate performance of a particular kind, the greater the chances that the exercise will fall outside of respondents' experiences" (Lloyd-Jones, 1977, p. 42).

Decisions about what kind of topic to use interact with the choice of rating procedures used to judge the writing samples. Many testing agencies use the method of holistic rating called General Impression Marking, in which the rater fits a writing sample into an ordered ranking on the basis of the total impression created by the paper. This is the procedure developed and used by the Educational Testing Service. It allows for wide-open topics, and treats writing in different discourse modes alike. However, as Lloyd-Jones (1977) points out, General Impression Marking reflects an assumption "that excellence in one sample of one mode of writing predicts excellence in other modes." He argues for a holistic rating system adapted to the kind of discourse being evaluated. Accordingly, he, Cooper (1977), Odell (1980, 1981) and others advocate selecting topics that elicit particular modes of discourse and designing rating guides which reflect the writer's performance on the special requirements of these topics. Lloyd-Jones (1977) has developed a procedure called "Primary Trait Scoring," which gives raters a scoring guide carefully adapted to the given topic. For example, his scoring guide for a persuasive topic on women's rights looks at such features as whether the writer takes a clear position, whether reasons are given for taking the position, what kind of reasons are given, and whether there is supporting elaboration.

The two views on rating procedures outlined here impose different criteria for the acceptability of topics. To those who consider wide-open topics unreasonable, the face validity of tests that employ them is diminished, regardless of the reliability of the test results. Thus it is possible to disagree on the validity of any given test, depending upon its design. Assuming that it is possible to agree upon what constitutes a valid topic and a valid design for a test, the next question to consider is whether, within that study or test, the ratings themselves will be assigned in a valid way.

Issues of Rating Validity: Readers, Criteria of Judgment, and Training

Consider the second condition for valid and reliable holistic ratings—that the readers are qualified and come from similar backgrounds. McColly (1970) stresses that even before they receive training in the holistic procedure, the readers who are chosen must be "competent." He defines competence in terms of scholarship and knowledgeability, and he draws a direct, though inferential, link between competence and validity: "The more competent the judges of essays are, the more they will agree and the more valid will be their judgments." Cooper (1977), citing the studies of Follman and Anderson (1967), further specifies that the readers must come from similar academic backgrounds so that they will draw as much as possible from common experience and values. He argues that a homogeneous group of raters can achieve high levels of reliability, as measured by statistical tests. Since we want readers whose judgments we would trust, these seem completely reasonable conditions. The difficulty arises when the selection of readers is considered against the conditions for selecting criteria of judgment, e.g., the rating guides, and the conditions for training readers to conform to these criteria.

There are two common ways to arrive at criteria of judgment. The first procedure is consistent with General Impression Marking. In this procedure, the criteria are arrived at inductively, by either the test-organizers or the readers, and may never be explicitly stated. The defining characteristic of this approach is that it weighs sample papers against each other, rather than against a pre-determined set of criteria. For example, McColly (1970) recommends simply selecting a group of readers whose common academic background creates a good chance for agreement, and allowing them to arrive at their own set of criteria. A similar procedure is described by Myers (1980), who suggests that the test-organizers select a set of "anchor papers" representing the range of papers produced at the test session. In order to select these papers, the test-organizers make a number of decisions, including what kind of paper to assign to each category; how many categories, or gradations in quality, to represent; and the range of quality to allow within a category. It is only after consensus is reached on the set of anchor papers that the organizers inductively define the characteristics that distinguish the categories. Readers are then trained to match samples to the anchors, although the categories may be modified to satisfy the opinions of the group. In any case, the initial selection of anchor papers by the test-makers crucially affects the evolution of the criteria.

The second common procedure for arriving at criteria is to formulate them in advance of the rating session. Rather than leaving readers to come up with their own set of criteria, the test makers develop rating guides or an explicit set of criteria to supply to the readers. Primary Trait Scoring is one

holistic procedure that uses explicit criteria. Scoring guides are drawn up for each topic, in accordance with the rhetorical situation it presents to the writer. In order to draw up the scoring guide, the test-makers have to decide what kind of writing the topic will elicit, what characteristics are necessary to that kind of writing, and what features of the writing samples will count as instances of those characteristics. Readers are then trained to judge the samples against the criteria.

Whether the criteria are implicit or explicit, they reflect the standards of the test-makers, the set of readers, or both. The decisions necessary to setting up categories or writing scoring guides are matters of opinion. Therefore, whatever criteria are evolved, they codify certain configurations of writing features as "best." However, we have seen that even readers who are well-qualified normally disagree about the worth of a piece. This means that while a set of criteria evolved by the test-makers, or by a homogeneous group of readers, is likely to satisfy that group, it is not guaranteed to satisfy writing experts at large. And yet, the standards of one group are imposed on all those who make use of the test results. A given set of criteria devised by one set of experts is no more valid than a different set of standards, arrived at by a different group of experts. Why should one set of criteria be imposed rather than another?

> [Diederich's work] has proved two significant facts: (1) that we disagree widely in our holistic judgments of writing, and (2) that the basis of our disagreements seems to lie in the different weights which we attach to a few traits of writing. This means that an inductive approach [to weighting the traits] cannot by itself lead to agreement, and that to impose an analytical weighting system based on the inductive results will actually work against widespread agreement. For, to use the categories about which readers *disagree* is to codify disagreement and to lose widespread acceptance from the start. (Hirsch, 1977, pp. 178–181)

Since, under normal circumstances, it is difficult to secure agreement on quality, any method that simply selects one standard is likely to be rejected by those who uphold alternative standards. As a result, the face validity of a given test of writing ability depends on whether one agrees with the criteria for judgment established for the ratings. For example, Vopat's (1982) rather strident charge that the Advanced Placement English Language and Composition Examination is invalid rests largely on objections to the criteria developed for the exam, as well as on objections to the selection of topics.

Whatever criteria are chosen, readers must be trained to use them. Training procedures are designed to "sensitize" the readers to the agreed upon criteria and guide them to employ those standards, rather than their own. The aspects of the training sessions that are intended to insure that readers use the right criteria are peer pressure, monitoring, and rating speed.

In one standard training procedure, readers practice by rating a set of essays written on the same topic as the test essays. The trainers then reveal any differences in the ratings and sometimes discuss them. Readers are expected to adjust themselves accordingly. Cooper (1977) believes that such peer pressure is effective. He cites Coffman's (1971) conclusion that: "In general, when made aware of discrepancies, teachers tend to move their own ratings in the direction of the average ratings of the group. Over a period of time, the ratings of the staff as a group tend to become more reliable." In order to maintain the influence of the group standard, Cooper recommends that the readers monitor themselves by periodically checking the reliability of their ratings during the actual scoring. Myers (1980) also recommends that readers monitor themselves. Monitoring by "table-leaders" is also a common practice. It is useful for detecting variance, caused in some cases by the onset of fatigue in the readers, which would reduce the statistical reliability of the results.

McColly (1970) emphasizes another factor: the speed with which readers rate the papers. He claims that with increased speed comes increased validity and reliability. "If a reader is competent, and if he has been well-trained and oriented, his instantaneous judgment is likely to be a genuine response to the thing for which he is looking. But if he is given time to deliberate, he is likely to accommodate his judgment to tangential or irrelevant qualities which will introduce bias into the judgment." McColly recommends monitoring the readers to keep them reading at a good speed, say one essay per minute for a 400-word essay. Similarly, Myers (1980) recommends instructing readers to "read fast, do not think about a paper too much and score your first impression, making certain it fits the anchors in front of you."

All of these training conditions tacitly imply that the readers' ability and willingness to conform to the agreed-upon criteria are extremely short-lived. Clearly, it must be difficult for readers to keep from applying their own idiosyncratic criteria. Readers must be trained not to apply them and must be monitored while they rate the essays because their adherence to the new criteria might slip over time; they may forget the criteria if they become tired. Readers must read quickly because on second thought their diverging, biased criteria will re-emerge. It seems that in order to achieve high reliability, testing agencies and researchers must impose a very unnatural reading environment, one which intentionally disallows thoughtful responses to the essays.

None of this would matter if it were certain that, in the end, the judgments were valid. Establishing the face validity of holistic ratings depends upon showing that the assessment is based on consistent application of acceptable criteria. We have seen that the selection of acceptable criteria is itself a difficult but necessary condition. The question now becomes whether trained readers actually adhere to whatever criteria are finally selected.

Adherence to Criteria: Research on Factors Influencing Holistic Ratings

Most of the empirical evidence available on holistic ratings concerns reliability. However a number of studies have correlated various characteristics of the rated essays with the holistic ratings they received. These studies indicate that, in spite of training, readers' judgments are strongly influenced by salient, though superficial, characteristics of the writing samples. Holistic ratings may produce high statistical reliability largely because they depend on characteristics in the essays which are easy to pick out but which are irrelevant to "true writing ability."

One of these superficial characteristics is physical appearance. McColly (1970) reports that the appearance of the writing sample (e.g., the quality of the handwriting) is strongly related to the holistic rating the paper will receive. Papers written in poor handwriting tend to receive lower holistic scores. McColly points out that when writing samples are rated in handwritten rather than typed form, the reliability coefficients of the ratings turn out spuriously high. People tend to agree on whether handwriting is good or not. Therefore, in a test that produces high statistical reliability, a significant amount of the agreement on the quality of the papers could be due to agreement on the appearance of the papers.

Word choice is another characteristic that can predict holistic ratings. Nold and Freedman (1977), Neilson and Piche (1981), and Grobe (1981) all found that the presence of uncommon or "mature" vocabulary items was strongly related to holistic ratings. Other factors that consistently correlate with holistic ratings are length of essay and spelling errors. The studies differ over the importance of syntactic maturity to high ratings. Nold and Freedman found evidence that the presence of final free modifiers, as in cumulative sentences, leads to higher ratings. On the other hand, Neilson and Piche failed to find evidence that complex headed nominals contribute to higher scores.

It is disconcerting to find holistic scores, which are supposed to be a *qualitative* measure, so directly predictable by such mundane quantitative measures as the length of the sample, the number of errors and the number of unusual vocabulary items. In fact, holistic scores are also correlated with sheerly quantitative "objective" tests. Godshalk et al. (1966) attempted to predict the holistic scores on students' essays from their SAT scores. They found that holistic ratings of students' essays can be predicted (with correlations of .70 and better), by the students' scores on quantitative tests, such as the English Composition Test, or the verbal sections of the SAT. They conclude that the quantitative tests must be valid because they assume that the writing sample scores are valid. Culpepper and Ramsdell (1982) also found positive correlations (ranging from .52 to .74) between a holistically scored essay test that they had developed and various quantitative measures, in-

cluding ACT scores, SAT scores, and an objective test which they devised. Stiggins (1982) reviewed five other studies conducted over the last six years and noted "a consistent and relatively strong correlation" between quantitative and qualitative tests.

The finding that holistic scores correlate with quantitative scores is interesting because quantitative tests have been so vehemently rejected as valid measures of writing ability. The fact that the quantitative and qualitative scores correlate does not establish that either one is valid, but merely that the two tests measure some of the same skills. The correlation can be interpreted in two ways. It might mean that the quantitative measures are more valid than they were thought to be. Or it might mean that the validity of the criterion, the holistic ratings, should be called into doubt. Godshalk et al. (1966) briefly consider the latter alternative: "One might argue that these findings are attributable to the fact that judgments based on short essays could reflect only the superficial, mechanical aspects of writing skill some critics claim as the functions measured by objective tests of writing ability." This possibility cannot be rejected lightly.

Insufficient research has been done into the question of whether readers trained in holistic rating base their judgments on substantive criteria or on superficial characteristics of the writing sample. Godshalk et al. (1966) cite a study by Myers, Coffman and McConville (1966) who claim that readers do make global judgments. Another step in this direction is the work of Freedman (1979). Freedman conducted an unusual experiment in which she manipulated the quality and quantity of certain characteristics of essays to be rated holistically. She manipulated the content, organization, sentence structure and mechanics in students' texts so that each of these features became either "strong" or "weak." For example, in essays with strong content features, all interpretations were sound, and all arguments were relevant, non-redundant, logically consistent, clear, and fully developed. In versions with weak content features, problems cropped up in each of those areas. Similarly, strong organization included proper paragraphing, logical order of presentation, and appropriate transitions. Strong sentence structure included mature and varied syntax and appropriate use of tense and reference. Strong use of mechanics reflected the standard rules for punctuation and spelling. Freedman found that although mechanics and sentence structure influenced the holistic scores, the content features predicted the scores best. Although significant, the results of Freedman's study cannot be applied directly to the validity question for several reasons. In order to manipulate the characteristics of the essays, Freedman rewrote them extensively. As a result, the essays that were tested reflected qualitative extremes rather than a natural distribution of writing abilities. In addition, the design of the experiment was not completely orthogonal: three-quarters rather than one-half of the

rewritten essays had strong content features. This means that only some of the necessary comparisons could be made.

Freedman's results do not conclusively confirm the validity of holistic scoring. On the other hand, other studies which showed a link between holistic scores and superficial features do not conclusively disconfirm it. Thus, the answer to the validity question is not a simple yes or no. Clearly, readers who assign holistic scores do not ignore the substantive features of the writing. It is unlikely that a high score would be assigned to a nonsensical paper solely on the basis of length, beautiful handwriting, and mature diction. What does emerge from these studies is a new set of questions. Are the readers predisposed by superficial features to be harsh or lenient in their application of substantive criteria? Do the reading conditions necessitated by holistic scoring increase the likelihood of such a predispostion?

The results of the correlational studies have an important consequence for those arguing for the predictive validity of holistic ratings: there is at present no "tried and true" criterion by which to establish such a claim. The goal in such studies is to establish the validity of a new measure by correlating its results with the results of a previously validated measure. Since the validity of quantitative methods as measures of writing ability is disputed, they cannot be used as the criterion. Many researchers, including Myers (1980) and Godshalk et al. (1966), have pointed out that the grades students earn in college English classes do not exclusively reflect their writing ability. So grades may not be an acceptable criterion either. Finally, since the validity of holistic scores is itself the point at issue, it would beg the question to correlate one set of holistic scores against another. And yet, these are the criteria by which ETS attempted to establish the validity of the Advanced Placement English Language and Composition Examination (Modu and Wimmers, 1981). Modu and Wimmers, in an article on ETS's results, compared the scores of Advanced Placement candidates on objective questions and holistically scored essays with the scores of college freshman English students who were given shortened versions of the same tests. Such comparisons can reveal differences between the two groups of students taking the test but don't provide conclusive evidence that the Advanced Placement exam is a valid measure of writing ability.

Conclusion: The Ultimate Usefulness of Holistic Ratings

Early attempts at qualitative evaluation of writing samples were abandoned because they were unreliable, not because they were invalid. However, the widespread confidence in the validity of current qualitative assessments must surely be tempered by considering the method of obtaining those assessments. Not any qualitative method will automatically be valid, even if it produces

reliable results. A writing sample may yet be the best, most valid representation of a writer's abilities. This paper has called the current method for evaluating writing samples into question. Is holistic rating a valid procedure for evaluating the writing samples?

The assumption that training in holistic rating leads readers to employ a consistent standard based on substantive criteria is not confirmed by the available evidence. In fact, there is evidence that holistic ratings may be reliable because, given the unnatural reading environment imposed upon the readers, the scores can only reflect agreement on salient but superficial features of the writing, such as the quality of the handwriting or the presence of spelling errors. To the extent that holistic ratings are intended to reflect substantive skills beyond the mastery of writing conventions, they must not be unduly influenced by superficial features if they are to be considered valid.

Perhaps more important than the effect of superficial features on readers is the finding that the criteria selected for judging the pieces of writing are themselves a matter of controversy. Holistic rating requires that readers be trained to use some consistent set of criteria, but at present these criteria have only *ad hoc* validity; they may be acceptable only to the group that formulates them. The choice of criteria affects both the reliability of the readers' judgments and the choice of topics. Since there is, at present, no agreement on how these choices are to be made, is it even possible to arrive at a generally accepted, valid set of criteria? Two approaches have been taken to this question, one by E. D. Hirsch, Jr., and the other by Richard Lloyd-Jones.

Hirsch (1977) believes in an abstract notion of good writing: "We cannot get reliable, independent agreement in the scoring of writing samples unless we also get widespread agreement about the qualities of good writing." Consistent with this belief, Hirsch attempts to develop a universal criterion of good writing, namely, "relative readability." Relative readability is a standard which measures how well ideas are presented. Since he believes that a writer's ability to present ideas remains fairly constant across writing tasks, Hirsch claims that reliable holistic scoring can be based on valid criteria: standards for relative readability. Hirsch has yet to accomplish this. In order for his approach to resolve the problem of choosing valid criteria, not only must Hirsch define the standards of relative readability and substantiate the face validity of these standards as a measure of writing ability, but he must also gain widespread acceptance for the standards.

In complete contrast to Hirsch, Lloyd-Jones (1977) believes that "good writing" cannot exist in the abstract. One can only judge whether or not a writing sample is a good example of a particular type of writing. Accordingly, Lloyd-Jones has developed a special kind of holistic rating, Primary Trait Scoring. The steps of his system are very carefully specified. Anyone designing a study which uses the system is required "to define the universe of discourse, to devise exercises [topics] which sample that universe precisely,

to ensure the cooperation of the writers, to devise workable scoring guides [created specifically for a particular topic] and to use the guides." Lloyd-Jones recognizes that there is a problem of validity in both the creation of criteria (scoring guides) and the application of the criteria by the readers. He claims that Primary Trait Scoring has advantages over other kinds of holistic scoring procedures, in that it requires extra care in eliciting and judging performance on particular kinds of discourse. Since the scoring guides are open to inspection, their validity can be considered in the public forum. If the methodology of Primary Trait Scoring is used as a research tool, it might be possible to define genres of writing tasks and arrive at generally accepted criteria for judging instances of these genres. Again, the bulk of the necessary work remains to be done.

The issue of criteria is central to the problems of holistic ratings as a valid means to evaluate writing ability. Settling this issue will require public discussion of the notion "good writing." It is not clear that this issue can be settled satisfactorily, but the approaches of Hirsch and Lloyd-Jones offer two places to begin the discussion.

A systematic exploration of the issue of criteria is necessary before we can be confident about the validity of qualitative evaluations of writing samples. In the meantime, there are ways to increase our trust in holistic ratings. First, we need more research along the lines of Freedman's (1979) study to determine how strongly such features as handwriting and spelling influence scores. Second, on a more practical level, we should take steps wherever possible to reduce the effect of such superficial features. For example, McColly (1970) recommends typing all papers before they are rated, to remove the effects of handwriting and neatness. (Unfortunately, such steps are expensive in terms of time and money, especially for large scale testing purposes.) Finally, whether scoring criteria are decided by readers or developed by test makers, they should be made available for public scrutiny. Since the validity of any given set of criteria is arguable, researchers and testers who employ holistic ratings ought to make the basis of their ratings clear.

The general questions that have been raised here about the validity of holistic ratings come in the face of increasing reliance of schools, researchers and testing agencies on this method of assessment. Holistic ratings should not be ruled out as a method of evaluating writing ability, but those who use such ratings must seriously consider the question of the validity of the scores that result.

References

Clifford, J. (1981). Composing in stages: The effects of a collaborative pedagogy. *Research in the Teaching of English, 15*, 37–54.

Coffman, W. (1971). On the reliability of ratings of essay examinations in English. *Research in the Teaching of English, 7,* 356-71.

Cooper, C. (1977). Holistic evaluation of writing. In Charles Cooper & Lee Odell (eds.), *Evaluating Writing.* Urbana, IL: National Council of Teachers of English. Culpepper, M. & Ramsdell, R. (1982). A comparison of a multiple choice and an essay test of writing skills. *Research in the Teaching of English, 16,* 295-297. Davis, K. (1979). Significant improvement in freshman composition as measured by impromptu essays: A large scale experiment. *Research in the Teaching of English, 13,* 45-48.

Diederich, P. (1974). *Measuring growth in English.* Urbana, IL: National Council of Teachers of English.

Flower, L. & Hayes, J. R. (In press). Process-based evaluation of writing: Changing the performance not the product. To appear in Douglas Buttruff (ed.), *The Psychology of Composition.* Conway, AK: L & S Books.

Follman, J. & Anderson, J. (1967). An investigation of the reliability of five procedures for grading English themes. *Research in the Teaching of English, 1,* 190-200.

Freedman, S. (1981). Influences on evaluation of expository essays: Beyond the text. *Research in the Teaching of English, 15,* 245-255.

Freedman, S. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology, 71,* 328-338.

Gere, A. (1980). Written composition: Toward a theory of evaluation. *College English, 42,* 44-58.

Godshalk, F., Swineford, F. & Coffman, W. (1966). *The measurement of writing ability.* Princeton: English Testing Service.

Grobe, C. (1981). Syntactic maturity, mechanics and vocabulary as predictors of quality ratings. *Research in the Teaching of English, 15,* 75-86.

Hilgers, T. (1980). Training college composition students in the use of freewriting and problem-solving heuristics for rhetorical invention. *Research in the Teaching of English, 14,* 293-307.

Hirsch, E. D., Jr. (1977). *The philosophy of composition.* Chicago: University of Chicago Press.

Lloyd-Jones, R. (1977). Primary trait scoring. In Charles Cooper and Lee Odell (eds.), *Evaluating Writing.* Urbana, IL: National Council of Teachers of English.

McColly, W. (1970). What does educational research say about the judging of writing ability? *Journal of Educational Research, 64,* 147-156.

Modu, C. & Wimmers, E. (1981). The validity of the advanced placement English language and composition examination. *College English, 43,* 609-620.

Moslemi, M. (1975). The grading of creative writing essays. *Research in the Teaching of English, 9,* 154-161.

Myers, A., Coffman, W. & McConville, C. (1966). Simplex structure in the grading of essay tests. *Educational and Psychological Measurement.*

Myers, M. (1980) *A procedure for writing assessment and holistic scoring.* Urbana, IL: National Council of Teachers of English and Educational Resources Information Center.

Neilson, L. & Piche, G. (1981). The influence of headed nominal complexity and lexical choice on teachers' evaluation of writing. *Research in the Teaching of English, 15,* 65-74.

Nold, E. (In press). Revising. To appear in C. Frederikson, M. Whiteman & J. Dominic (eds.). *Writing: The nature, development and teaching of written communication.* Hillsdale, NJ: Lawrence Erlbaum.

Nold, E. & Freedman, S. (1977). An analysis of readers' responses to essays. *Research in the Teaching of English, 11,* 164-174.

Odell, L. (1981). Defining and assessing competence in writing. In Charles Cooper (ed.), *The nature and measurement of competency in English.* Urbana, IL: National Council of Teachers of English.

Odell, L. & Cooper, C. (1980). Procedures for evaluating writing: Assumptions and needed research. *College English, 42,* 35-43.

Sanders, S. & Littlefield, J. (1975). Perhaps test essays can reflect significant improvement in freshman composition. *Research in the Teaching of English, 9,* 145-153.

Shuy, R. (1981). A holistic view of language. *Research in the Teaching of English, 15,* 101-112.

Stiggins, R. (1982). A comparison of direct and indirect writing assessment methods. *Research in the Teaching of English, 16,* 101-114.

Vopat, J. (1982). Comment and response. *College English, 44,* 532-539.

Announcement and Call for Papers Penn State Conference on Composition and Rhetoric

Wayne Booth, Peter Elbow, and James Kinneavy will be the major consultants at the third Penn State Conference on Rhetoric and Composition to be held July 10-13, 1984, at State College, Pennsylvania.

People interested in participating are invited to present papers, demonstrations, or workshops on topics related to rhetoric or the teaching of writing—on composition, rhetorical theory and history, basic writing, technical and business communication, advanced composition, and so forth. One-page proposals will be accepted until April 15.

If you wish to submit a proposal or volunteer to chair a session, or if you are interested in more information about attending or participating in the conference, write to Professor Jack Selzer, Department of English, The Pennsylvania State University, University Park, PA 16802