Evaluation: A Holistic Perspective

Pamela Russell
Université de Sherbrooke

In holistic evaluation, the evaluator views the text as a whole, determining the degree to which it is effective as a specific type of writing. This paper examines the principles, applications, and limitations of holistic evaluation, and explores the contextualization of holistic evaluation standards. A number of holistic scoring guides are described; the differences between them reveal that different genres exhibit different features in their successful forms. The paper also describes an empirical study into the correlation between analytic and holistic evaluations of a corpus of summaries written by university-level students.

THIS PAPER WILL EXPLORE WAYS IN WHICH RESEARCH into holistic evaluation can contribute to our understanding of the bases on which we judge writing. I will begin by giving an overview of holistic evaluation and examining what a holistic procedure of evaluation is, what the literature reports on it, and how it is applied in the assessment of writing skills. In so doing, I will deal with an issue that is not adequately addressed in the literature: namely, the contextualization of the standards which underlie a holistic assessment. The second part of this paper will describe the results of an empirical study I conducted into the quality of summaries, using holistic evaluation as a research tool. The results of this study illuminates the relationship between analytic and holistic evaluation, and between tacit and expressed criteria in evaluation.

This paper thus addresses the concerns of teachers of professional and technical writing, who must evaluate a variety of genres of writing according to a wide range of context-specific criteria. When such criteria are contrasted, it becomes apparent that different genres exhibit different features in their successful forms. In particular, the description of criteria used in evaluating summaries (a standard assignment in a professional writing course) should prove to be of interest to the writing instructor.

Defining Holistic Evaluation

To begin, let me define my terms. By evaluation, I refer to the procedure by which one assesses the quality of a text and attributes some sort of a score or rating to it. By "holistic" evaluation, I refer to the process by which an evaluator assesses a text according to his or her impression of its overall merit, rather than by isolating and measuring specific features. The term is used inconsistently in the literature on the subject: it is used to refer to everything from the vaguest of impressionistic approaches, to a very structured procedure in which standards of assessment are closely described. In all cases, however, it refers to a vision of the text as a whole rather than as a sum of its parts.

Basically, holistic evaluation is one pole of a continuum, the opposite pole of which is analytic evaluation. Most teachers use neither a purely holistic nor a purely analytic approach; rather, they combine approaches, trying to reach an assessment that takes into account the complex interplay of individual features, their cumulative effect, and their importance in a particular writing context.

Charles Cooper uses the term "holistic evaluation" in the broadest of ways, to include "any procedure which stops short of enumerating linguistic, rhetorical, or informational features of a piece of writing" (1977, p. 66). He considers what he calls "general impression marking" to be one variety of holistic scoring, and includes in his examples of holistic procedures approaches that others might well classify as analytic.

In *Creating Writers*, Spandel and Stiggins differentiate between two types of holistic evaluation, which they call "general impression holistic scoring" and "focused holistic scoring" (1990, p. 6). In focused holistic scoring, writing samples are matched against set standards and scored accordingly. (What these standards are and how they are context-specific are issues to be addressed later in this paper.) In the field of language testing, numerous empirical studies of such focused holistic scoring have been carried out. These studies usually involve tests that are rated by a group of evaluators; the evaluators' ratings are then averaged for each paper. Such studies emphasize the importance of evaluators' determining in advance common standards that define each of the possible ratings. In general, the procedure is as follows: a number of possible ratings are set (e.g., from 1 to 5) and each rating is defined in terms of specific criteria. The most common way of defining criteria is by means of a scoring guide, which describes the qualities of a text that merits each specific rating. Standards can also be set indirectly, through representative papers. In this case, evaluators are shown samples of writing, called "anchor papers" (Spandel and Stiggins, 1990, p. 20), which are representative of each rating. Once the standards have been set, evaluators read through the full set of papers, matching each one with the scoring guide description or with the anchor papers they judge to be of similar quality.

Validity and Reliability of Holistic Evaluation

Perhaps the best known and most damning study of holistic evaluation was carried out by Paul Diederich (1974), and is described in his book *Measuring Growth in English.* Diederich selected a corpus of 300 compositions written by college-level students and had them evaluated by 60 readers from six different fields. His readers included English teachers, science teachers, editors, lawyers, and business people. The readers were instructed to sort the papers into nine piles in order of general merit, and they were left on their own to decide what constituted general merit. The results showed very little agreement among readers: ratings were very inconsistent, and the interrater reliability measured was only .31: a very low score, for Diederich considers an acceptable reliability rating to be .80 . Diederich concluded that his readers had judged the papers on very different bases, and as an alternative to the holistic approach, he suggested an analytic procedure: he analyzed text quality into specific features to be evaluated separately. The features he isolated for evaluation fall into two general categories: ideas, organization, wording, and flavour (which he termed "general merit"); and usage, punctuation, spelling and handwriting (which he termed "mechanics"). The total rating would then be determined by the sum of the scores for the individual features.

It is not surprising, however, that Diederich's experiment with holistic evaluation produced such negative results. Consider the diversity of readers he used. With such a wide range of readers, all judging according to their own views of what good writing is, the results were bound to be inconsistent. Moreover, he gave his readers no direction, and left them to assess papers on the basis of their undirected reactions. Yet holistic scoring need not be so haphazard and arbitrary. It is far more successful when it is focused, through scoring guides or anchor papers.

A question that arises is thus: is holistic evaluation valid? Does it measure what it is intended to measure? In focused holistic evaluation, the answer is yes, since the evaluators work with standards that define what they are attempting to measure, and what a successful paper is in a given situation. The contextualization of those standards is of paramount importance, particularly for the professional writing teacher, who must evaluate a variety of genres such as compositions, journalistic-style articles, business reports, sets of technical instructions, and summaries.

In holistic evaluation, rather than judging a text on the basis of isolated features, the evaluator judges to what extent it fulfills its purpose. Cooper writes, "Holistic evaluation is obviously to be preferred [to analytic evaluation] where the primary concern is with evaluating the communicative effectiveness of candidates' writing" (1977, p. 3). Yet these notions of "purpose" and "communicative effectiveness" are too vague to provide concrete guidance for the evaluator. For holistic evaluation to produce valid results, the evaluator needs to define these notions in terms of context-specific criteria, and then focus on the cumulative effect of a text's specific features on the communication process.

Another question that arises is that of reliability: is holistic evaluation reliable? Does holistic evaluation produce consistent results? Inter- rater reliability is generally measured by having different markers score a set of papers and then determining the average correlation between their scores. There have been numerous studies of reliability, many of which are reported in Cyril Weir's *Communicative Language Testing* (1990). The findings of these studies have varied. In general, they have indicated that focused holistic scoring produces acceptable reliability scores, comparable to those achieved by analytic procedures.

Holistic evaluation has proven to be most successful when practiced by experienced evaluators; it is more difficult for those who are new and relatively inexperienced. The evaluator must not only monitor a wide range of features at one time but he or she must also contextualize them and assess their cumulative effect rather than simply categorizing and weighing them individually, as in an analytic evaluation.

Holistic scoring is particularly appropriate when a group of papers needs to be rated on a continuum, or when a group of students needs to be rank-ordered—for example, in placement tests. On a practical level, a real advantage to holistic scoring is its speed—it is certainly much faster for an experienced evaluator to read through a corpus of texts and sort them according to overall merit than to identify and categorize all the errors and then calculate the mark on the basis of individual performance of subskills. This advantage is, of course, somewhat offset by the fact that holistic evaluation is much more reliable when papers are evaluated by more than one evaluator.

A purely analytic approach is unsatisfactory for a number of reasons. One major problem is that it is difficult to delimit the subskills that constitute writing competence. E.M. White points out the limitations of analytic scoring in his book *Teaching and Assessing Writing*:

> In theory, analytic scoring should provide the diagnostic information that holistic scoring fails to provide and in the process yield a desirable increase in information from the writing sample. In practice, three major problems have so far demonstrated the limitations of analytic scoring: (1) There is as yet no agreement (except among the uninformed) about what, if any, separable subskills exist in writing. (2) It is extremely difficult to obtain reliable analytic scores, since there is so little professional consensus about subskills. (3) Analytic scoring tends to be quite complicated for readers. (1985, pp. 29-30)
>
> There is no evidence that writing quality is the result of the accumulation of a series of subskills. To the contrary, the lack of agreement on subskills in the profession suggests that writing remains more than the sum of its parts and that the analytic theory that seeks to define and add up the subskills is fundamentally flawed. (p. 123)

Scoring Guides

As we have seen, scoring guides are often used to ensure the reliability of holistic scoring when it is carried out by a group of evaluators. Scoring guides are also useful for the individual teacher: by drawing up such a guide, the teacher can define the qualities that define an A, B, C, D or F paper for a specific writing activity. Thus the teacher-evaluator reflects on and establishes standards in his or her mind, in light of the purpose and genre of the writing activity, the communicative context, and the level of the class. But because terms such as "purpose" are rather vague, concrete examples of scoring guides are needed to illustrate how evaluation criteria are context-specific. Let us therefore look at excerpts from a number of scoring guides to see the variety of criteria by which different types of writing are assessed.

In the teacher's manual that accompanies *Reporting for the Print Media*, author Fred Fedler (1989) suggests criteria for grading journalistic articles—

in other words, he gives a scoring guide for journalism teachers. He defines an A paper as follows:

> The story is newsworthy and exceptionally well written: thorough and free of errors. The lead is clear, concise and interesting . . . . The body is well organized and contains effective transitions, quotations, descriptions, and anecdotes. Because of the story's . . . merit, newspapers would . . . publish it. (p. 2)

He begins his description of a B paper as follows: "The story could be published by a newspaper after minimal editing"(p. 2). And an F paper is described thusly:

> The news story could not be published by a newspaper, nor easily rewritten. It is too confusing, incomplete or inaccurate. Or, the story contains a misspelled name or serious factual error. (p. 2)

The last criterion for an F paper—the presence of a misspelled name or factual error—reflects the fact that in journalism, where one writes in a public forum, the cardinal sin is getting a name or fact wrong, since such a mistake could ruin reputations and lead to costly libel suits.

Obviously, very different criteria would be used to evaluate the writing of ESL students. One description of a top-ranking paper in ESL is as follows: "The writing is indistinguishable from that of a native speaker." Similarly, a colleague who specializes in ESL showed me a scoring guide she uses, from which I take the following definition of a B paper:

> Student writes clearly understandable English and organizes material well. Grammatical errors are . . . not serious enough to interfere with communication . . . . Sentence structure may be somewhat inelegant, but is clear and understandable. (G. Arbach, personal communication, June 10, 1990)

Criteria used by evaluators of ESL writing thus indicate that evaluators measure the students' ability against that of a native speaker, emphasizing comprehensibility and idiomaticity.

A scoring guide used by the British Council to determine university admission is based on nine ratings. The description of the top rating begins as follows: "The writing displays an ability to communicate in a way which gives the reader full satisfaction"; and the next rating begins with: "The writing displays an ability to communicate without causing the reader any difficulties" (Hughes, 1989, p. 88). The ratings are more fully described, but the criteria cited indicate that the texts are evaluated partly in terms of the reader's response.

So scoring guides describe in broad strokes the standards by which the merit of texts is to be determined in a particular writing activity. Their purpose is to ensure that impressionistic or intuitive assessment does not translate into arbitrary or idiosyncratic assessment. They also serve to make explicit the expectations that underlie the evaluator's intuitive response. No doubt teachers of professional writing, who assign a variety of writing tasks to their students, would be well advised to identify, for themselves and for their students, the features that characterize successful texts of different genres. For it is apparent that not only will different genres exhibit different features in their successful forms, but also that features crucial to success in one genre may be relatively unimportant in another.

Applications in Research: An Empirical Study

Holistic evaluation is a useful research tool to help elucidate those bases on which we evaluate writing. I used holistic evaluation for this purpose in an empirical study I conducted into the quality of summaries written by university-level students. In this study, a set of student texts was evaluated both holistically and analytically. I determined the correlation between the holistic scores, and specific analytic variables, in order to identify those variables that carried the most weight in a holistic evaluation, as well as the relationship between tacit and expressed criteria.

In brief, I had 55 students write summaries of a text taken from a magazine article on Costa Rica. I then submitted the summaries to four evaluators, all experienced university teachers of writing, who were instructed to rate each summary holistically on a scale of 1 to 6, with a score of 6 indicating an excellent summary and a score of 1 indicating a completely unacceptable summary. To ensure that standards and expectations were similar, I drew up a holistic scoring guide for this writing task, and consulted the evaluators about it; all agreed that the guide reflected our expectations of summary-writing quality in student summaries. (This scoring guide is given in the Appendix.)

The evaluators read the summaries quickly, and gave each one a rating from 1 to 6, based on their overall impression of the summary's merit. For each paper, I averaged the four scores to determine the holistic score—a procedure that has produced highly reliable results in other studies.

The results of the holistic evaluation were as follows. First, the interrater reliability was calculated to be approximately .72 (with a p-value of .0001). Perhaps one reason that the reliability was not higher was that one of the evaluators did not rate any of the papers as a 6—she did not think that any of them matched the description of an excellent paper. This points to a clear difference between two types of holistic evaluation: one in which texts are scored simply on a curve, on the basis of their relative merit, with the best texts in the group receiving the highest score; and the other, in contrast, a holistic evaluation in which texts are scored according to their absolute merit, on the basis of set standards. The evaluator who didn't give any of the texts a 6 was measuring the texts against a set standard, and was not simply rating them in relation to another.

In the second part of this study, I carried out a detailed analysis of each of the summaries, and rated each text according to eight variables: errors of grammar and mechanics, distortions of meaning, inclusion of important ideas, integration of important ideas, syntactic complexity of the sentences (measured by number of T-units per sentence and average length of T-unit), organization of important ideas, and efficiency of summarization. I then analyzed the correlation between the holistic scores determined in the first part of the study and the eight variables I had identified. The most significant correlation was between the holistic evaluation and three of the variables: first, and primarily, the inclusion of important ideas; secondly, the absence of errors of usage and grammar; and thirdly, the absence of distortion. A stepwise linear regression procedure carried out to determine the combination of variables that would best predict performance showed those three variables explained 76% of the variance of the holistic scores.

Particularly indicative of the complex role of error in evaluation was a comment made by one of the evaluators. In one of the summary texts, the student writer had consistently misspelled the name Costa Rica, calling the country Costo Rico. Now misspelling is usually viewed simply as a mechanical error, a surface error easy to identify and classify. In error analysis, what could possibly be more straightforward than a spelling mistake? Yet the evaluator saw this misspelling as something more serious. She commented, "This misspelling seriously affected my impression of the text's merit. After all, the whole text is about Costa Rica, and if the writer doesn't even get the name of the country right, how effective is the summary?" (K. Barber, personal communication, September 2, 1991).

This anecdote illustrates that, when viewed according to their effect on communication rather than error typology, spelling mistakes are not all equal. This same observation was apparent in Fedler's scoring guide, mentioned earlier, in which misspelling a person's name in a news story was viewed as an error of such importance that it could earn the writer a failing grade. In these cases, the type of error does not necessarily indicate its communicative effect.

Let's examine this anecdote in the context of the theoretical model I used for my research into summaries: Kintsch and Van Dijk's (1978) macropropositional representation of the meaning of a text. Kintsch and Van Dijk describe a processing model of discourse comprehension and production, and characterize a text's semantic structure on two levels: the microstructure and the macrostructure. The former term refers to the local structure of individual propositions and sequences of propositions, the latter, to the "gist" of the text—its global structure. Both structures are abstracted from the surface structure and are described in terms of sequences of propositions. Thus the macrostructure of a text is a hierarchical representation of its gist, or overall meaning.

In the case in question, the spelling error occurred at the highest level of the macrostructure. The importance placed on the misspelling of Costa Rica can be attributed to the fact that it was the topic of the text—the highest-level argument.

Thus the level of macrostructure on which an error occurs, along with the effect of an error on the reader, is an important consideration in error analysis.

It was a holistic approach to evaluation that brought to light the importance of this specific error, for the holistic approach focuses on the effect of text features in specific genres, and on the response the text elicits in the evaluator as reader. By being overly analytic and by focusing on types of errors rather than on their effects, an evaluator may fail to account for the complex interplay of form and function. One can conclude that the holistic dimension of text quality should be recognized and included in any evaluation procedure.

Conclusion

This paper has reviewed the concept and practice of holistic evaluation and has attempted to show that there are issues involved in holistic evaluation that remain unresolved and may be fruitfully explored—issues such as the determination of context-specific criteria, the description of features exhibited by different genres in their successful forms, and the analysis of the complex nature of error. Moreover, by illustrating the use of holistic evaluation as a research tool, I have tried to show how a holistic evaluation procedure can help us learn more about standards used both explicitly and implicitly to determine the quality of a text.

References

Bamberg, B. (1982). Multiple choice and holistic scores: what are they measuring? *College Composition and Communication, 33*, 404-406.

Cooper, C. C. (1977). Holistic evaluation of writing. In C.R. Cooper & L. Odell *(Eds.), Evaluating writing: describing, measuring, judging* (pp.3-29). Urbana, Ill.: NCTE.

Diederich, P. B. (1974). *Measuring growth in English.* Urbana, Ill.: NCTE.

Fedler, F. (1989). *Reporting for the print media*: *Instructor's Manual.* New York: Harcourt Brace Jovanovich.

Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press.

Kintsch, W., & Van Dijk T.A. (1978). Toward a model of discourse comprehension and production. *Psychological Review, 85*, 363-394.

Russell, P. (1992). *The integration of theory and practice in the development of summary-writing strategies.* Unpublished doctoral dissertation, University of Montreal.

Spandel, V. & Stiggins R.J. (1990). *Creating writers: linking assessment and writing instruction.* New York: Longman.

Van Dijk, T.A., & Kintsch W. (1983). *Strategies of discourse comprehension.* New York: Harcourt Brace Jovanovich.

Weir, C. (1990). *Communicative language testing.* Toronto: Prentice-Hall.

White, E. M. (1985). *Teaching and assessing writing.* San Francisco: Jossey-Bass Publishers.

Appendix

Proposed Holistic Scoring Guide For Summaries

6: The summary-writer demonstrates a clear understanding of the original and presents the ideas lucidly and coherently. The summary contains all of the major ideas of the original, without distortion, and demonstrates the writer's ability to distinguish between main ideas and secondary information. Ideas have been combined and integrated where appropriate. The text is virtually free from errors of mechanics, usage, and sentence structure.

5: The summary-writer demonstrates a clear understanding of the original and conveys the major ideas with little distortion. The text may contain minor weaknesses of structure or clarity, but demonstrates the writer's ability to present information coherently. Ideas have been largely combined and integrated where appropriate. The text is fairly free from serious errors of mechanics, usage, and sentence structure.

4: The summary-writer more or less adequately conveys the main ideas of the original, although there may be some weaknesses in the discrimination between major and secondary ideas, or in use of language. The summary may insufficiently develop certain ideas, or may contain some distortion, but it demonstrates the writer's basic comprehension of the original. There is evidence of some combination and integration of ideas where appropriate. The summary is organized and written well enough to allow the reader to comprehend reasonably easily, although it may be disjointed or lack focus in places. The summary may contain errors of grammar and usage, but not so frequently as to raise serious doubts about the writer's competence in English.

3: The summary-writer has some difficulty conveying the main ideas of the original. He or she may fail to discriminate between major and secondary ideas, may omit or distort major ideas, or may copy sections of text verbatim and fail to integrate them into the text. Errors in grammar, usage, and sentence structure may interfere with readability. Despite definite weaknesses in selection, development of ideas, or expression, the text is still intelligible.

2: The summary-writer does not adequately convey the major ideas of the original, because of omission, distortion, poor analysis, or inability to express ideas clearly. There is evidence of some or all of the following problems: errors in comprehension; lack of coherence between sections; or frequent errors in grammar, usage, and sentence structure. The general impression is that of confused thinking and poor writing.

1: The summary-writer fails to convey the major ideas of the original. This may be because of the writer's misunderstanding of the original, because of lack of organization and development, or because of an inability to write intelligibly.

The above scoring guide is inspired by and adapted from the holistic model for evaluating compositions given in White (1985, pp.135-36).