ED 315 799 CS 506 910

AUTHOR Goulden, Nancy Rost

TITLE Theoretical and Empirical Comparisons of Holistic and Analytic Scoring of Written and Spoken Discourse.

PUB DATE Nov 89

NOTE 27p.; Paper presented at the Annual Meeting of the Speech Communication Association (75th, San Francisco, CA, November 18-21, 1989).

PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142) -- Information Analyses (070)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Comparative Analysis; Higher Education; *Holistic Evaluation; *Rating Scales; Secondary Education; Speech Communication; Test Reliability; *Test Theory; Test Validity; Verbal Communication

IDENTIFIERS *Analytical Scoring

ABSTRACT

Since speech communication evaluators are beginning to adapt the analytic and holistic instruments and methods used for rating written products to oral products and performance, this research review investigated: (1) what the labels "analytic" and "holistic" mean; (2) the theoretical bases of the two scoring approaches; and (3) the rather limited experimental studies both in written and oral discourse which compare the two methods of using rating scales. The paper's six sections are as follows: (1) Understanding the Terms; (2) Use of Terms in Writing and Speaking Literature; (3) Theoretical Bases for Holistic and Analytical Scoring; (4) Holistic and Analytic Scoring and Purpose of Evaluation; (5) Reliability and Validity of Holistic and Analytic Scoring; and (6) Empirical Studies of Reliability and Validity of Two Methods. Sixty-one references are attached. (SR)

Theoretical and Empirical Comparisons of Holistic and Analytic Scoring of Written and Spoken Discourse

Nancy Rost Goulden
Speech Department
Kansas State University
Manhattan, Kansas 66506
(913) 532-6875

Presented at the Speech Communication Association: San Francisco, CA: November, 1989

In direct opposition to the current vogue in educational assessment in general, the Assessment Section of the summer 1988 SCA Communication Education Conference went on record as endorsing the view that the focus of communication evaluation should be the classroom. "The best hope for assessment is the classroom teacher who is qualified in communication in an instructional setting where communication principles and skills are taught. It would be better use of our time to empower the classroom teacher" (p. 1). Stiggins (1988) further points out "we are again failing to address the central issue in school assessment: insuring the quality and appropriate use of teacher-directed assessments of student achievement used every day in classrooms from coast to coast" (p. 363).

The search for the best means to reach the goal of "quality" "teacher-directed assessments" of oral communication products, processes and performances has been extensive and long standing. Much of the ongoing discussion about oral communication assessment has centered on the development and refinement of rating scales and their use by educators to score public speeches (Applbaum, 1974; Becker, 1962, 1965; Clevenger, 1963, 1964; Stevens, 1928; Thompson, 1943, 1944; Wiseman and Barker, 1965).

Oral Communication assessment in 1989 obviously encompasses student performance in a variety of communication contexts, some more complex than the speaker-audience format; however much classroom assessment in speech communication continues to be product/performance evaluation utilizing rating scales. Therefore assessment issues arising from rating scales and how they are used to rate human discourse continues to be an important aspect of the search for fair and useful evaluation by the classroom teacher.

For the most part, past research of speech rating scales has focused on the building of valid scales, training of raters to use the scales accurately while avoiding undue rater error (Applbaum, 1974; Backlund, 1983; Becker, 1970; Bock and Bock, 1982; Bohn and Bohn, 1985; Ragsdale, 1972). Recently investigators have started to reconsider an issue raised by Thompson in 1944: does the format of the scale and the method that format represents have an effect on the rater's score?

In the area of assessment of written composition, the types of rating scale instruments and subsequently the methods of scoring written products (essays and compositions) have for the most part become standardized into two categories: analytic scoring methods and holistic scoring methods. As speech communication evaluators begin to adapt the instruments and methods used for rating written products to oral products and performance, it is necessary to investigate (1) what the labels "analytic" and "holistic" mean, (2) the theoretical bases of the two scoring approaches, and (3) the rather limited experimental studies both in written and oral discourse which compare the two methods of using rating scales.

Understanding the Terms

"Holistic Scoring" is defined as a method of evaluation which considered the overall quality of the product/performance, includes the component parts and traits of the object of evaluation but does not mark them separately or add subscores for a composite score. Exploring the features of holistic scoring and clarifying the differences which separate holistic scoring from analytical scoring, atomistic scoring, and general impression scoring will provide a base for looking at the literature specific to methods of scoring.

Even though definitions for holistic evaluation vary among contemporary researchers, there does seem to be agreement that when raters are grading holistically, they will react to the entire speech or composition rather than only respond to isolated parts or features. However this response to the entire product is not the primary delimiting feature of holistic scoring which separates it from analytic scoring. For when using the analytic approach, a rater may also attempt to evaluate the entire product. Lloyd-Jones (1977) uses the term "atomistic" as the opposite of holistic to refer to "assessments of particular features associated with skill in discoursing, whereas holistic tests consider sample of discourse only as whole entities" (p. 33). For holistic scoring all major features are factored into the evaluation. The scoring is not based on just delivery skills of the speaker or the level of creativity of the writer but the impression of those discrete traits when combined with all other traits contributing to the whole product.

It is also important to clarify the difference between holistic scoring and general impression scoring. General impression scoring has the same features as holistic scoring except that the criteria are usually chosen by the individual and may not be articulated. The scoring that classroom teachers carry out independently when they read, comment on and then assign a grade to an essay is considered general impression scoring by Quellmalz (1982) and White (1985). In holistic scoring a more formal and group-centered process leads to selection and recording of criteria, followed by deliberate effort by all raters to base the scoring on the same set of criteria.

Most of the time when evaluators of written discourse refer to holistic scoring, they expect that a general set of criteria designed to encompass all modes of writing will be used. A special sub-category of holistic scoring called "focused-holistic scoring technique" by McCready and Melton (1981) in their review of a number of writing evaluation techniques from a group of state assessments uses the elements of holisticism with a mode-specific rubric. For example, an educator designs criteria specifically to score a persuasive speech for a "focused-holistic" rating scale which includes items not found in a general public speaking scale.

When the term holistic scoring is understood, it becomes relatively easy then to define the contrasting method of analytic scoring. In analytic scoring, evaluators may end up judging the whole, but only by first analyzing the separate parts and then combining that series of judgments. It is this recording and combining of subscores which separates analytic and holistic scoring. Scherer's (1985) explanation of the use of an analytical scale included the essential elements of analytic scoring, rating individual components of the whole product/performance and adding the subscores for an overall score of the product/performance. Individual scales may be weighted. Some analytic instruments include a scale labeled "general impression" which is added into the composite score.

Use of Terms in Writing and Speaking Literature

In the materials consulted for this study, at least 19 different definitions for holistic scoring in the evaluation of writing were utilized. The term was originally used to describe large-scale, standardized scoring of writing samples administered either by testing agencies or educational organizations such as school districts or institutions of higher learning. The holistic scores from these evaluations are used for placement, admissions, or to fulfill competency testing requirements, all purposes which call for an overall evaluation of writing.

The specific method most closely associated with the term holistic scoring is the rather rigid system of scoring writing samples developed by and prescribed by Educational Testing Service. This process consists of six steps designed to produce maximum reliability between and within raters (O'Donnell, 1984; Fowles, 1978). For wide-scale testing, written compositions are often ranked relative to other products being evaluated at that time (O'Donnell, 1984). However, essays scored holistically may be either ranked or rated or both (Myers, 1980).

When White (1985), Freedman and Calfee, (1983), McCready and Melton, (1981) and Quellmalz, (1982) write about holistic scoring of writing, their definitions do not limit the method to the policies appropriate for wide-scale testing only, but are based on an understanding which allows flexibility of procedure to make holistic scoring usable in a variety of classrooms.

In speech communication, writers have usually named the methods of scoring by the labels of types of rating scales. Thompson (1944) tackled the problem of the effect of how we rate in the early forties by comparing the scoring of the same speeches using different instruments which represented analytic, holistic, and general impression scoring. Here are his equivalent terms: "descriptive scale" = holistic; "linear, attitudinal, and diagnostic scales" = analytic; "letter grades" = general impression.

In the sixties and seventies, several researchers did comparative studies using what were called "simple" ("general effectiveness") and "complex" scales (Clevenger, 1964; Applbaum, 1974). The simple scale is a general impression scale, producing a single global score. The complex scale is a bit of hybrid. In the Clevenger and Applbaum studies, raters scored a number of discrete traits, concluding with a separate scale for "general effectiveness." Instead of comparing a sum of the trait scores with or without the general effectiveness score, the researchers compared the simple scale score with only the "general effectiveness" score from the complex scale. First scoring single traits is an analytic approach, but failing to sum or failing to require that the overall trait correlate with the single trait is a holistic approach. Therefore, using complex scales in these studies seems to represent a combined analytic-holistic method of scoring.

The Wiseman and Barker (1965) comparative investigation of methods of rating speeches recognizes that the scale itself is not the determina. t of the method, that one scale can be used in more than one way. Subsequently, these researchers seem to be describing distinct analytic and holistic methods of scoring when they write about "composite ratings" (analytic totals) and "overall ratings" (holistic judgment) utilizing the same rating instrument. During the present decade some speech evaluators have begun to use the term holistic scoring. Taylor (1987) conducted a holistic speech-rating study, based on the definitions of holistic scoring from the writing field. When speaking of system-wide oral placement evaluation, Willmington (1986) used the term "single holistic rating" (p. 6). These two examples suggest that the borrowed term from writing evaluation is becoming a part of the speech communication evaluation vocabulary.

Theoretical Bases for Holistic and Analytical Scoring

The underlying rationales for the holistic and analytical methods demonstrate a kind of antithetical stress expressed in the answers to the following questions: (1) Is the whole more than the sum of the parts? (2) Do the parts generate the whole? (3) Does the whole generate the parts?

Although based on writing evaluation, White's (1985) response to these questions fits equally well for speech evaluation. "Holisticism in grading reflects the view that writing activities are not describable through an inventory of their parts" since the "holistic approach denies that the whole is only the sum of its parts" (p. 18).

White's (1985) corollary is that humans are also more than the "sum of their behaviors" (p. 15). An important goal of education is to teach students the necessary behaviors for socialization. In discourse these are such behaviors as correct usage in language presentations, standard pronunciation, patterns of discourse organization. Many of the entries on discourse rating forms reflect the socialization traits in speaking and writing. If educators only graded these behaviors, the task in performance/product evaluation might be easier. But as White (1985) points out what makes a human more than the sum of his behaviors and what makes an essay or speech more than the sum of its parts is the antithesis of socialization, individualization.

Therefore since raters using a holistic approach are not limited to only the specific entries on a rating sheet or the exact descriptions of a rubric, they can respond to individual expression which reflects the individuality of the author/presenter. In holistic scoring, the rater has some jurisdiction to weigh in the unexpected features students include to place their own individual stamp on their discourse.

The analytic approach for scoring speeches and essays in contrast is based on the premise that the whole is the sum of the parts, which are "neatly sequential and comfortably segmented" (White, 1985, p. 30). In an attempt to make the scoring of a product more objective, reliable and at the same time more valid by reflecting both socialization and individual traits, a proponent of an analytic approach might try to design the ultimate, complete rating scale, but such a project would almost certainly be doomed to failure. Lloyd-Jones (1977) projects that "it may be simply that the categorizable parts are too numerous and too complexly related to permit a valid report" (p. 36). This is not an indictment of analytic scores but merely an explanation of the contrasting bases for holistic and analytical scoring.

In harmony with the view that scoring and summing certain individual aspects of a performance or product can never be enough for a valid grade, holistic raters do not give subscores or focus on "separable aspects" (White, 1985) but give one overall score on the whole product based on a general impression (O'Donnell, 1984; Freedman and Calfee, 1983; Scherer, 1985; Fowles, 1978).

Just because holistic raters do not mark the "parts" does not mean that holistic scorers ignore the components of a speech or writing product. McCready and Melton (1981) suggest that when grading writing compositions holistically, the rater realistically cannot be oblivious to the elements that make up the whole. "For holistic scoring, a general impression of the student's ability to use the mechanics correctly is a consideration in assigning a score. A similar technique is applied in analytic scoring except that a separate score for mechanics is assigned" (p.79). There is another important distinction between the parts as they are considered holistically and analytically in addition to whether the rater puts down a symbol for each trait. In analytic scoring the exact trait is chosen and described in precise language and that description forms the boundary that the rater must stay within. In holistic scoring the rubric is usually a more general description of trait categories at different levels of effectiveness. As indicated above the rater may add traits not found in the guide. In holistic scoring, the whole generates the parts (Hake, 1986) rather than the parts forming together to produce the whole.

Hake (1986) in writing about discourse evaluation borrows from Piaget's model, "a whole (which generates its parts) has external relationships that influence the internal parts" (p. 156). This statement is related to the expectations and standards which each rater brings to the rating task. Raters have a conceptualization of the whole product and what it should and shouldn't contain.

Hake's view suggests that when scoring holistically, we focus on that concept and see where the product we are evaluating falls short or exceeds our expectation. This is why general impression scoring results are so varied: the parts generated by varied concepts of the whole product/performance may be exceedingly different. Holistic rubrics attempt to describe at least the framework of a consistent whole for the raters so they will generate consistent parts and then be able to score based on the presence, absence, and level of those parts while maintaining the flexibility to add other parts that reflect the individuality of the product.

Comparison of Holistic and Analytic Scoring

Holistic and Analytic Scoring and Purpose of Evaluation

The differing theoretical bases are one way of comparing and contrasting the two methods of using a scoring guide. Returning again to the criteria of fair and useful grading, it may be assumed that the underlying principles of a method could be a consideration in determining if that method is fair and useful. However, the best theoretical fit may depend on the match of the method to the purpose of a given evaluation. The purpose of some evaluations is to assess the whole; others to assess the parts. If the purpose is to sort students or assign an end of term grade, then the method might be the summative method of holistic marking (Cooper and Odell, 1977). However, if the purpose is to give an analysis of strengths and weakness of a diagnostic speech or meaningful feedback to be used by the student to work for improvement, an analytic evaluation might be more appropriate.

There is no hard fast rule that precludes using holistic and analytic technique in concert. For example, both McCready and Melton (1981) and Quellmalz (1982) describe situations where papers were first scored holistically, then analytically for feedback.

Reliability and Validity of Holistic and Analytic Scoring

Discourse evaluation literature points out it is not enough that the method fit the assessment situation. It must be established that either or both methods are reliable and valid when used for evaluating speeches if instructors are to meet the criteria of fairness to students in grading. Only then can a recommendation be made for one or the other or both methods.

One might think of methods of rating products/performances as a continuum of decreasing reliability as the technique moves from atomistic to analytic (Becker, 1970) to holistic to general impression. Because atomistic scoring usually just involves counting or noting presence or absence of elements, the results from this method can be replicated by other raters or the same rater at another time (Lloyd-Jones, 1977). Each method as the chain progresses from atomistic to general impression calls for an increased degree of individual subjective judgment which would logically suggest less consistency among individuals.

It might be expected that a general validity continuum also exists moving in the opposite direction through the same sequence from highest validity for general impression scoring to lowest validity for atomistic scoring. Depending on such a pattern is even more suspect when talking about validity than when looking at reliability. Except for instrument reliability, reliability of rating is determined by analyzing the raw scores at the end of the process. Validity is often established by comparison with outside subjective judgments. Simply because there are more subjective variables both inside and outside, it is more difficult to pin down validity and easier to claim that validity exists. Consequently, the position of a method relative to other methods is difficult, perhaps impossible to establish. As a result, a comparison of validity for analytic and holistic scoring depends to an extent on the logical arguments of supporters of each method.

For example, Wiseman and Barker (1965) argued that analytic scoring of speeches may reduce rater bias since the raters are less aware of what their subscores mean and therefore less apt to force the score to fit their preconceptions. In the writing literature, Cooper, (1977, p. 3) champions the validity of holistic scoring, "holistic evaluation of writing remains the most valid and direct means of rank-ordering students by writing ability" (p. 3). The claim for the validity of holistic scoring in writing is also supported by Lloyd-Jones (1970, p. 33) when he states that holistic scoring is "potentially more valid." The holistic argument discussed above, the whole is greater than the sum of its parts, is additional defense for holistic scoring being a more valid method than a method that can never include all the parts.

When questioning the validity of much of the holistic scoring currently carried out in writing assessment, Charney (1984) goes back to the difficulty of establishing validity when she writes of the dual problems of validity: the criteria or rubric have to be valid and the application of that scoring guide by the raters has to be valid. The first step to determine if what Ragsdale (1972) calls "the problem of validity....fidelity to a stated objective" (p. 334) has been achieved is to look at interrater and intrarater reliability. But as Backlund (1983) reiterates, "Reliability is necessary for validity, but it does not guarantee it" (p. 67).

It is clear there must be some other means to try to establish if the method of rating is producing a true or valid evaluation. Both in writing and in speaking, concurrent validity is established by comparing scores from an experiment with scores from a panel of experts (Quellmalz, 1982; ETS Quality Assurance Free Response Testing Team, 1987; Thompson, 1944; Gundersen, 1978; Marzano, 1975).

A second method of establishing concurrent validity is to compare the raters' scoring for each speaker or writer to some other measure that purports to measure the same competencies. Willmington (1986) compared the scores for speakers from placement performance given near the end of the semester with scores from four speeches given during the semester in a public speaking class by those speakers and found a reliability coefficient of .82, suggesting that if the speech grades were valid, the placement performance was also. This suggests that the course grade if it is primarily based on performances may be used to establish concurrent validity (Willmington, 1983).

Empirical Studies of Reliability and Validity of Two Methods

Empirical information about reliability and validity of holistic and analytic scoring seems to fit into three categories (1) conclusions in summarizing articles which refer to empirical studies, (2) investigations in which data related to only one method of scoring was collected, and (3) studies which compared more than one method of rating.

In an overview of writing assessment, Quellmalz (1982) documents the statement "When carefully structured scale training sessions precede actual rating, most holistic and analytical rating scales can demonstrate high interrater reliability" (p. 11) with five sources from 1979 and 1980. Much of the popularity of holistic scoring in writing assessment is the result of the high reliability of the raters' scores as reported in a study by Fred I. Godshalk, Frances Seinford, and William E. Coffman which investigated the reliability of the writing component of the College Entrance Examination Board now under ETS (Bauer, 1982; White 1985).

In 1970 Becker concluded, based on the experimental work in speech evaluation to that time, that a general impression scale preceded by scoring subscales was more reliable than general impression scoring alone. He also suggested that true analytical scoring is more reliable than an overall score. This assertion is based on an unpublished study by Becker (1962).

More recent studies (from category 2 above) using only holistic scoring for speaking have confirmed that sound reliability is possibie for speech evaluation. Willmington (1986) states that although reliability varied on speech assessments for placement at several campuses of the University of Wisconsin, the coefficient for rating was as high as .89 at one site. Taylor (1987) translated the holistic scorin technique ured for writing to the rating of speeches and found satisfactory reliability. Five judges "were within a one-point agreement range on 84% of their ratings and within two points of each other (using a four-point scale) on 99% of the ratings" (p. 4).

Unlike the speech studies of Willmington and Taylor, Marzano's study of the scoring of writing samples compared methods of scoring (category 3). Marzano (1975) looked at the validity of analytic and holistic methods by comparing scores on essays to criteria scores produced by experts' rating of the essays. He found the holistic method produced higher validity (.80) than the analytic method (.47). When the experiment scores were compared for reliability, the analytic approach proved to be more reliable (.70 versus .59) (p. 4).

Bauer's comparative study, also of writing samples, focused on reliability. Bauer (1982) collected data to describe interrater reliability and intrarater reliability for nine raters' scores on writing samples generated by three different rating methods: analytic, primary trait and holistic. She found satisfactory reliability for all three methods on both varieties of reliability; however, the reliability was slightly higher for analytic and primary trait scoring than for holistic scoring. The procedure for implementing the methods were those already established for wide-scale scoring in the writing field. No new information was collected about validity by Bauer.

Turning again to speech evaluation, three experimental studies focus directly on the effect of method on reliability. In Clevenger's 1963 study the primary emphasis was on retest reliability by having raters score filmed speeches a second time after an interval of six weeks. The method of scoring was the combined holistic-analytic system where raters score independent traits first and then give a general effectiveness score. Although these raters were not scoring using two distinctly different methods at two different times, the analysis of the separate scales provides some comparative information. Earlier studies had led Clevenger to expect that reliability would be higher on the discrete scales (which are closer the analytical scoring) than on the more global entry (which is more like holistic scoring). However, he reports these results: "Just the reverse seems to be the case in this experiment; for the three most reliable scales include the two most global or general traits, while the five least reliable scales include the four traits which are generally assumed to represent relatively uncomplicated, directly observable speech characteristics" (p.

290).

The early studies of Thompson (1944) in speech evaluation attempt to determine if the method of scoring has an effect on the scores given for speeches. However Thompson used a variety of different instruments, each representing a method of scoring, rather than comparing two methods using one basic set of criteria. His results may have been influenced as much by the instruments as by the method; however, the instruments were used to compare primarily analytic scoring with general impression scoring and holistic scoring with general impression scoring. An experiment comparing analytic and holistic methods was not included. Nevertheless, from the comparisons he did make, Thompson concluded, "The differences among rating methods are less than is believed. Whatever advantages there are in accuracy (reliability) favor the simple devices" (p. 78).

Of the experiments described in the speech communication literature, the work of Wiseman and Barker (1965) is the most closely allied to this proposed study since in the Wiseman Barker study one instrument was used by the same raters to score speeches using two different methods analogous to holistic and analytic scoring. The raters were a group of students enrolled in public speaking classes and a group of instructors of public speaking classes. Reliability related to method varied between the two groups. Instructors rated consistently whether they used an analytic or a holistic approach. When student raters' scores were the result of the sum of the subscores (analytic), they were consistent with instructors' ratings; however when students rated holistically, their "overall ratings were consistently higher than instructor evaluations" (p. 135). Although final examination and final course grade were available, the researchers did not consider the question of concurrent validity. However, the comparison of student rater's evaluation of speeches to those of "expert" instructor raters, could be used to make the argument that based on this form of concurrent validity, analytic scoring in this study was more valid. Wiseman and Barker do not use the term validity in discussing the results.

In the quest for "quality" evaluation of oral communication in the classroom, there is a continuing need for educators to design appropriate instruments and score oral products and performances by valid and reliable methods. The writing community has been exploring analytic and holistic scoring both through academic discussion and empirical experiments for over ten years now. The speech communication community is just beginning to be aware that although we have used both methods, separately and in combination, almost from the time we became a separate discipline (Stevens, 1928), we are only just beginning to look intently at the theoretical underpinnings of the two approaches and to test the reliability and validity of the two methods for assessing the progress and level of achievement of our students in oral communication classrooms.

References

Applbaum, R.L. (1974). Intrarater reliability: A function of scale complexity and rater training. Central States Speech Journal, 25, 277-281.

Allen, R.R., Wilmington, S.C., and Sprague, J. (1972). Speech communication in the secondary school. Boston: Allyn and Bacon.

Assessment resolution #1. (1988) Speech Communication Association Summer Conference. Flagstaff, AZ.

Backlund, P. (1983). Methods of assessing speaking and listening skills. In R.B. Rubin (Ed.), Improving speaking and listening skills (pp. 59-73). San Francisco: Jossey-Bass.

Bauer, B.A. (1982). The reliabilities and the cost-efficiencies of three methods of assessment for writing ability and empirical inquiry. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Becker, S.L. (1962). The rating of speeches; scale independence. Speech Monographs, 29, 38-44.

Becker, S.L. (1970). Rating scales. In P. Emmert and W.D. Brooks (Eds.), Methods of research in communication (pp. 213-235). New York: Houghton-Mifflin.

Becker, S.L. and Cronkhite, G.L. (1965). Reliability as a function of utilized scale steps. The Speech Teacher, 4, 291-93.

Bock, D.G. and Bock, H.E. (1982). Evaluating classroom speaking. Annandale, VA: Speech Communication Association.

Bock, D.G. and Munro, M.E. (1979). The effects of organization, need for order, sex of the source, and sex of the rater on the organization trait error. The Southern Speech Journal, 44, 364-72.

Bock, D.G. and Saine, T.J. (1975). The impact of source credibility, attitude valence, and task sensitization on trait error in speech evaluation. *Speech Monographs*, *42*, 229-36.

Bohn, C.A. and Bohn, E. (1985). Reliability of raters: The effects of rating errors on the speech rating process. *Communication Education*, *34*, 343-351.

Brooks, K. (1957). The construction and testing of a forced choice scale for measuring speaker achievement. *Speech Monographs*, *24*, 73.

Brossell, G. (1986). Current research and unanswered questions. In K.L. Greenberg, H.R. Wiener, and R.A. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 168-182). New York: Longman.

Burry, J. and Quellmalz, E. (1983). *Assessing student's writing skills: The CSE expository and narrative writing scales*. Los Angeles: University of California.

Charney, D.C. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English*, *18*, 65-81.

Clevenger, T., Jr. (1963). Retest reliabilities of 10 scales of public speaking performance. *Central States Speech Journal* *14*, 285-91.

Clevenger, T., Jr. (1964). Influence of scale complexity on the reliability of rating general effectiveness in public speaking. *Speech Monographs*, *31*, 153-56.

Conlan, G. (1985). *How the essay in the college board English composition test is scored: An introduction to the reading of readers*. Princeton, NJ: Education Testing Service.

Cooper, C.R. (1977). The holistic evaluation of writing. In C. Cooper and L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 3-32). Urbana, IL: National Council of Teachers of English.

Cooper, C.R. and Odell, L. (1977). *Evaluating writing: Describing, measuring, judging*. Urbana, IL: National Council of Teachers of English.

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

Diederich, Paul B. (1977). Definition of ratings on the ETS composition scale. (ERIC Document Service No. ED 145-454).

ETS Quality Assurance Free-Response Testing Team. Breland, H., Conlan, G., Fowles, M. (chair), and Livingston, S. (1987). Guidelines for developing and scoring free-response tests. Princeton, NJ: Educational Testing Service.

Fowles, M.E. (1978). Manual for scoring the writing sample. Analytical scoring, holistic scoring. Princeton, NJ: Educational Testing Service.

Freedman, S.W. and Calfee, R.C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, and S.A. Walmsley (Eds.), Research in writing: principles and methods, (pp. 75-98). New York: Longman.

Gay, L.R. (1981). Educational research: Competencies for analysis and application. Columbus, OH: Charles E. Merill Publishing Co.

Geyerman, C.B. and Bock, D.G. (1984, November). The effects of dogmatism, rhetorical sensitivity and attitude valance on selected speech ratings. Paper presented at the meeting of the Speech Communication Association, Chicago, IL.

Gundersen, D.F. (1978). Video tape modules as a device for training speech raters. The Southern Communication Journal, 43, 395-406

Hake, R. (1986). How do we judge what they write? In K.L. Greenberg, H.R. Wiener, and R.A. Donovan (Eds.), Writing assessment: Issues and strategies, (pp. 153-167). New York: Longman.

Houston, J.E. (Ed.), (1986). Thesaurus of ERIC descriptors. Phoenix, AZ: Ongx Press.

Lederman, M.J. (1986). Why test? In K.L. Greenberg, H.R. Wiener, and R.A. Donovan (Eds.), Writing Assessment: Issues and strategies (pp. 35-43). New York: Longman.

Lloyd-Jones, R. (1977). Primary trait scoring. In C.R. Cooper and L. Odell (Eds.), Evaluating Writing: Describing, measuring, judging (pp. 33-68). Urbana, IL: National Council of Teachers of English.

Marzano, R.J. (1975). On the validity of analytic ratings. Unpublished manuscript, University of Colorado, Denver. (ERIC Document Reproductions Service No. ED 112 412)

McCready, M.A. and Melton, V.S. (1981). *Feasibility of assessing writing using multiple assessment techniques research report*. Ruston, LA: Louisiana Technical University. (ERIC Document Reproductions Service No. ED 220 871)

Mitchell, K. and Anderson, J. (1986). Reliability of holistic scoring for MCAT essay. *Education and Psychology Measurement*, *46*, 771-775.

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

O'Donnell, H. (1984). Large scale writing assessment. *ERIC Digest*. Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills. (ERIC Document Reproductions Service No. ED 250 691)

Quellmalz, E.S. (1982). Designing writing assessments: Balancing fairness, utility and cost. Los Angeles: California University Center for the Study of Evaluation. (ERIC Document Reproductions Service No. ED 228-270)

Quellmalz, E.S. and Capell, F. (1979). Defining writing domains: Effects of discourse and response mode. Los Angeles: California University Center for the Study of Evaluation. (ERIC Document Reproductions Service No. ED 212 661)

Ragsdale, J.D. (1972). Evaluation of performance. In W.W. Braden (Ed.), *Speech methods and resources* (pp. 420-440). New York: Harper and Row.

Scherer, D.L. (1985). Measuring the measurements: A study of evaluation of writing: An annotated bibliography. Bloomington, IN: Indiana University. (ERIC Document Reproduction Service No. ED 260 455)

Steele, J.M. (1985, March). Trends and patterns in writing assessment. Paper presented at 3rd Annual Conference on the Assessment of Writing. San Francisco, CA.

Steele, J.M. (1986). *The assessment of reasoning and communicating (ARC) norms and interrelationships.* Unpublished manuscript.

Steele, J.M. (1987). College outcome measures program. Iowa City, IA: College Testing Program.

Stevens, W.E. (1928). A rating scale for public speakers. The Quarterly Journal of Speech, 14, 223-232.

Stiggins, R.J. (1988). Revitalizing classroom assessment: the highest instructional priority. The Kappan, 69, 363-368.

Stiggins, R.J. and Bridgeford, N.J. (1984). The nature, rate, and quality of performance assessment in the classroom. Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA.

Tamor, L. and Bond, J.T. (1983). Text analysis: Infering process from project. In P. Mosenthal, L. Tamor, and S.A. Walmsley (Eds.), Research on writing: principles and methods (pp. 99-138). New York: Longman.

Taylor, K.P. (1987, November). Holistic evaluation of speaking skills. Paper presented at meetings of Speech Communication Association. Boston, MA.

Thompson, W.N. (1943). Is there a yardstick for measuring speaking skill? Quarterly Journal of Speech, 29, 87-91.

Thompson, W.N. (1944). An experimental study of the accuracy of typical speech rating techniques. Speech Monographs, 11, 65-77.

Tinsley, H.E.A. and Weise, D. (1975). Interrater reliability and agreement of subjective judgements. Journal of Counseling Psychology, 22, 358-376.

Trank, D.M. (1983, April). Assessing growth in a communication skills program. Paper presented at meetings of Central States Speech Association. Lincoln, NE.

White, E.M. (1985). Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performances. San Francisco: Jossey-Bass, Inc.

Willmington, S.C. (1983). Assessing oral communication performance skills. Paper presented at meetings of Speech Communication Association. Washington, D.C.

Willmington, S.C. (1986). The development of performance tests for a speech placement program. Paper presented at the annual meetings of the Central States Basic Course Director's Conference. Stillwater, OK.

Wilson, F.R. and Griswold, M.L. (1985). The effects of method and comprehensiveness of training on the reliability and validity of ratings of counselor emphathy. *Measurement and Evaluation in Counseling and Development*, *18*, 3-11.

Wilson, L.R., Scherbarth, B.D., Brickell, H.M., Mayor, S.T., and Paul, R.H. (1988). *Determining validity and reliability of locally developed assessments 1988*. Flossmoor, IL: Illinois State Board of Education.

Wiseman, G. and Barker, T.L. (1965). A study of peer group evaluation. *Southern Speech Journal*, *31*, 132-38.

Witte, S.P., Trachsel, M., and Walters, K. (1986). Literacy and the direct assessment of writing: A diachronic perspective. In K.L. Greenberg, H.S. Wiener, and R.A. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 13-34). New York: Longman.