

BRUTAL CHOICES IN CURRICULAR DESIGN ...

HOLISTIC SCORING

BY WILLA WOLCOTT

Willia Wolcott teaches at the University of Florida in Gainesville and serves as an associate chief reader for the holistic scoring of several state-mandated tests. She has published articles and a book on writing assessment.

Brutal Choices in Curricular Design ... is a regular feature of Perspectives, designed to explore the difficult curricular decisions that teachers of legal research and writing courses are often forced to make in light of the realities of limited budgets, time, personnel, and other resources. Readers are invited to comment on the opinions expressed in this column and to suggest other "brutal choices" that should be considered in future issues. Please submit material to Helene Shapo, Northwestern University School of Law, 357 East Chicago Avenue, Chicago, IL 60611, phone: (312) 503-8454, fax: (312) 503-2035.

Prologue

Holistic scoring, though most commonly associated with national or statewide writing assessments, offers definite potential for legal writing classes as well. For example, it can provide legal writing teachers an alternate way of grading selected written assignments. Not only are teachers experienced in holistic scoring able to grade large numbers of essays efficiently, but the very use of this scoring approach also forces teachers to evaluate global, as well as surface, features of the essays. Training new teachers in holistic scoring procedures can thus help to prevent novice instructors from overattending to mechanics or to any single writing trait in their grading procedures. In addition, as previously reported in *Perspectives*, proposed changes to the Law School Admission Test (LSAT) include the possible addition of a scored writing assessment, designed to emphasize critical and analytical writing skills.¹ Having familiarity with holistic scoring procedures should, therefore, be beneficial to legal writing instructors.

Law students themselves can benefit from having some of their assignments holistically

¹ Annemarie Bridy, *A New Direction in Writing Assessment for the LSAT*, 11 *Perspectives: Teaching Legal Res. & Writing* 61 (2003).

scored.² For example, holistic scoring guides enable students to see the breadth of criteria on which overall writing judgments are based and to engage in classroom discussions about what good writing entails. Although holistic scoring does not provide specific, diagnostic feedback about individual writing traits the way that analytic scoring does, students can, nevertheless, see from the descriptions contained in scoring guides the typical qualities embodied by the scores their own work has received. Students may then use these guides (which they sometimes help to develop) to practice rating either a series of sample essays or their own writing drafts.

Nature of Holistic Scoring

Based on the theory that the whole is more than the sum of its parts,³ holistic scoring emphasizes the need for having scorers evaluate an essay in terms of its overall impression. The impression is not a snap judgment; rather, it is derived from the readers' thorough understanding of the criteria and their training in applying those criteria to papers. Considering all writing elements without focusing unduly on any single trait, scorers make no marks on an essay; instead, they read each paper quickly, and then, without rereading it, assign a single score on the basis of how successfully various writing traits, such as development, focus, clarity, organization, diction, and mechanics, combine to work together within a piece.

Background

Holistic scoring has been greatly influenced by the classic work of Paul Diederich, John French, and Sydell Carlton,⁴ who sought to determine grading standards and the factors that accounted for grading differences among essay readers. They asked 53 professionals from six diverse fields each to rank order in nine separate piles according to quality about 50 papers (300 papers in all) written

² For a discussion of the use of holistic scoring in classrooms, see M.L. Rodgers, *How Holistic Scoring Kept Writing Alive in Chemistry*, 43(1) *College Teaching*, 19–22 (1995).

³ Miles Myers, *A Procedure for Writing Assessment and Holistic Scoring* (1980).

⁴ *Factors in Judgments of Writing Ability*. Research Bulletin RB-61-16, 1961.

“Based on the theory that the whole is more than the sum of its parts, holistic scoring emphasizes the need for having scorers evaluate an essay in terms of its overall impression.”

“Holistic scoring emphasizes the importance of building—collegially—group consensus as to the standards adopted...”

at home by college freshmen from different universities. No training or guidance was provided the readers. A large percentage of the essays (94 percent) received at least seven of the nine possible grades, and no paper was given fewer than five grades. Differences among readers fell into five broad areas: ideas, form, flavor, mechanics, and wording. The authors stressed both the need for general consensus among readers and the importance of providing them with training and direction if scorings were to serve major purposes.

A second influential study was *The Measurement of Writing Ability* by Fred Godshalk, Frances Swineford, and William Coffman.⁵ The authors used holistic scoring to evaluate each of five essays written by nearly 650 high school juniors and seniors throughout the country. Although their study was undertaken primarily to validate multiple-choice items, they concluded with several recommendations for holistic scoring procedures.

Other studies explored the elements that evaluators focus upon when making judgments about writing quality. Some researchers found readers to be strongly influenced by organization and content,⁶ whereas others found readers to be greatly concerned with such elements as mechanics, spelling, vocabulary, or length.⁷ Still others concluded that readers' expectations about the writers can influence results.⁸ However, it is important to note that, as Brian Huot cautions, several studies used methodologies other than

traditional holistic scoring procedures, a factor which might have affected the findings.⁹

Training Procedures

Procedures currently used for formal holistic scorings often follow those established by the Educational Training Service. Holistic scorers are typically trained to evaluate essays according to established criteria contained in a scoring guide, which conveys the qualities a generic essay reflects at each score point for a particular exam.

These guides may be derived either descriptively or prescriptively. That is, in some assessments the guides have been descriptively written based on qualities reflected in papers actually scored on two or three occasions. In other assessments, the scoring guides are prescriptively written beforehand to reflect the qualities expected at each scoring level.

In a scoring guide, the criterion for a trait such as diction may be generically described as “sophisticated and precise” in a top-scoring paper, whereas it may be labeled “pedestrian” or “general” in a lower-half paper. Because an actual essay rarely matches the entire criteria described for each score point, sample papers on current topics are essential for training purposes. The sample papers are preselected from actual assessment essays by chief readers and then validated by scoring leaders prior to a scoring session in order to illustrate how the scoring guide can be applied at each scoring point to papers on a given topic.

Holistic scoring emphasizes the importance of building—collegially—group consensus as to the standards adopted, and both the training and monitoring procedures are designed to facilitate this group consensus. At the start of a scoring session, readers are reminded of the testing conditions under which students wrote their essays and to keep certain principles in mind.¹⁰ Readers are urged to a) disregard the significance of length per se as some short papers may be compact and

⁵ ETS Research Monograph No. 6 (1966).

⁶ See Sarah Freedman, *How Characteristics of Student Essays Influence Teachers' Evaluations*, 71(3) *Journal of Educational Psychology* 328–338 (1979); see also Hunter Breland & Robert Jones, 1(1) *Perceptions of Writing Skills*, 1(1) *Written Communication* 101–119 (1984).

⁷ See Cary Grobe, *Syntactic Maturity, Mechanics, and Vocabulary as Predictors of Quality Ratings*, 15(1) *Research in the Teaching of English* 75–85 (1981). See also Bennett Rafoth and Donald Rubin, *The Impact of Content and Mechanics on Judgments of Writing Quality*, 1(4) *Written Communication* 446–458 (1984). See Carl Stach, *The Component Parts of General Impressions: Predicting Holistic Scores in College-Level Essays*, 1987 *Dissertation Abstracts International*, 48, 07A. (University Microfilms No. DES 87-22706).

⁸ See Loren Barritt, Patricia Stock & Francelia Clark, *Researching Practice: Evaluating Assessment Essays*, 37 *College Composition and Communication* 315–327 (1986).

⁹ See *The Influence of Holistic Scoring Procedures on Reading and Rating Student Essays*, in *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations* 206–236 (Michael Williamson and Brian Huot, eds. 1993).

¹⁰ For a good listing of procedural scoring criteria, see *A Guide to the New SAT Essay* 12 (2004).

succinct, whereas long papers may ramble or collapse, b) remain alert to any papers that may trigger their own biases, and c) turn papers with difficult handwriting or other problems over to scoring leaders for review. Then scorers begin their training by rank ordering from best to worst a set of preselected sample papers known as rangefinders. The rangefinders serve as “anchor papers,” or papers that reflect each scoring level on the currently assigned topic. As the set of papers contains at least one anchor paper for each point on the scoring scale, the rank ordering procedure enables scorers to see the quality of essays in relation to one another. This concept of rank ordering is important, for it means, as Edward White points out, that a given score is not an absolute.¹¹ Rather, it must be understood in the context of the entire scale against which the essays are ranked. Thus, unlike a classroom grade of B or A, a paper cannot be given a holistic score—a three, for instance—out of context. Once scorers have rank ordered the papers, in centralized scoring sessions they are asked to reveal publicly the scores they have assigned. This public tallying of scores enables scorers to see how closely their scores correspond to those given by other readers. (Online scoring has taken the place of centralized scoring sessions for some assessments; however, readers still are trained and monitored by scoring leaders online.)

Throughout the scoring session, readers are subsequently given at regular intervals pairs of additional prescored sample papers that they, in turn, must grade to ensure they are continuing to uphold group standards. At this point, readers are not explicitly rank ordering papers; rather, they are assigning scores to papers by employing the criteria they have internalized for each scoring point. Their performance is further monitored by two additional methods—first, by scoring leaders who randomly select papers from each reader for review and second, by check readings, in which all participants—the chief readers, the scoring leaders, and the readers—score a batch of new essays to ensure they are scoring alike as much as possible. The emphasis placed on consistency of scores has

been criticized by some writing scholars, who question the validity of such readings, given the complexity of the writing process.¹² However, studies done by Brian Huot and later by Judith Pula and Huot to determine the issue of validity, rather than reliability alone, found that the very training procedures entailed in holistic scorings not only assisted readers in making their evaluations of essays but actually freed the scorers to engage in a fuller response to the essays they scored.¹³ Certainly, the group consensus established helps to minimize idiosyncratic preferences on the part of scorers, thereby increasing interrater reliability rates and ensuring more fairness for the writers. Scorers who are unwilling to adopt the group standards may be disqualified from participating.

Scorers and Scoring Scale

For many assessments, scorers are chosen with a strong writing background in their professional field. Essays are generally scored by two readers who assign coded scores independently of one another. Although the breadth of the scale used varies according to the purpose of the assessment, a scale of six points is typical in that it allows for some discrimination among scoring levels without requiring readers to make too fine a distinction. Further, most scales used contain an even number

“Essays are generally scored by two readers who assign coded scores independently of one another.”

¹² Davida Charney, *The Validity of Using Holistic Scoring to Evaluate Writing: A Critical Overview*, 18 *Research in the Teaching of English* 65 (1984). Charney states, “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.” *Id.* at 79. See also Peter Elbow, *Embracing Contraries: Explorations in Learning and Teaching* (1986). Like Charney, Elbow expresses reservations about an evaluation model that requires agreement among judges. Not only may it result in an overemphasis on surface features of grammar and mechanics, in his view, but also it requires readers to suspend their own judgments in favor of other standards. Similarly, William Smith points out that disagreements are bound to occur even among trained scorers, just as they occur among such other groups as “trained literature specialists” and “trained critics.” See Smith, *Assessing the Reliability and Adequacy of Using Holistic Scoring of Essays as a College Composition Placement Technique*, in *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations* 142, 198 (Michael Williamson and Brian Huot, eds. 1993).

¹³ *Supra* note 7. See also Judith Pula and Brian Huot, *A Model of Background Influences on Holistic Raters*, in *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations* 237–265 (Michael Williamson and Brian Huot, eds. 1993).

¹¹ Edward White, *Teaching and Assessing Writing* (1985).

“Regardless of the number of score points, it is important to recognize that each score point represents a range along a continuum.”

of points so that readers do not settle arbitrarily for the middle point. Indeed, readers are urged to distinguish first in their own minds whether a paper is upper half or lower half before deciding on the actual level.

Regardless of the number of score points, it is important to recognize that each score point represents a range along a continuum. There can be a strong “six” paper, for example, or a “six” paper that is “looking down,” just as there can be a weak “two” or a solid “two” paper. Precisely because each score point signifies a range, contiguous scores from two readers are usually considered acceptable. For example, whereas one reader may see a given essay as a strong “four,” another reader may view it as a weak “five,” and they both may be right. If essays are scored two or more points apart, on most scales they are considered “splits”—that is, papers reflecting disagreement—and the essay is given to a third designated party such as the scoring leader to resolve. Splits may arise from a variety of factors: If a scorer becomes fatigued, the room is too warm, or other external conditions occur, then the scores a reader assigns may begin to drift upward or downward; conversely, if the papers themselves are problematic, such as showing a marked disparity in quality between content and surface features, then readers may tend to weigh these elements differently, and their scores may vary as a result. Although holistic scoring entails some subjectivity by the very nature of the human judgment involved, the extensive training and ongoing monitoring procedures are designed to minimize as much as possible the occurrence of splits and to maximize the consistency, and therefore the accuracy, of the scores that result. The value of the training and monitoring was noted by one scoring leader who participated in a holistic scoring study I previously conducted:

As a table leader, I have observed the monitoring process as a tempering of our individual prejudices and preconceived notions about how the papers should be graded. We must set aside our whims, caprices, and dogmatism in the interest of fairness and competency. Readers, table leaders, and chief readers balance papers

against group standards, adjusting skillfully as we proceed.¹⁴

New Scoring Developments

The increase of writing assessments at both state and national levels has focused new attention on computer-based or automated scoring. Different versions of computerized scoring are being developed, and several states have experimented with computerized essay scoring in pilot tests. According to some supporters, the technology offers the potential for increased cost-effectiveness, faster turn-around time, and a need for fewer human scorers.¹⁵ Advocates further note that computer evaluations are very reliable and that computers can agree more highly with human readers than pairs of readers do. However, as Donald Powers, Jill Burstein, Martin Chodorow, Mary Fowles, and Karen Kukich point out, the issue of validity is troublesome in that computers have been criticized for focusing on grammar, mechanics, and vocabulary.¹⁶ In an effort to see how well results obtained by one automated scoring system corresponded to other external writing criteria, Powers and his colleagues explored relationships between the scores that examinees for the GRE Writing Assessment received and non-test indicators, such as undergraduate writing samples, self-reports of writing grades, and self-evaluations of writing criteria. They found that the automated scores obtained by the automated scoring system e-rater related “modestly” to the other indicators of writing skill and in a “reasonably similar” manner to those obtained by human raters. They point to “the potential of automated scores as (valid) indicators of prospective graduate students’ writing skills, especially when they can be combined with scores provided by at least one human reader.”¹⁷

¹⁴ See Willa Wolcott, *Perspectives on Holistic Scoring: The Impact of Monitoring on Writing Evaluation*, 1989 Dissertation Abstracts International 51, 05A (University Microfilms No. DA9028582).

¹⁵ Andrew Trotter, *States Testing Computer-Scored Essays*, 21 (38) *Education Week*, May 29, 2002, at 1–2.

¹⁶ Donald Powers, Jill Burstein, Martin Chodorow, Mary Fowles & Karen Kukich, *Comparing the Validity of Automated and Human Scoring of Essays*, 26(4) *J. Educational Computing Research* 407 (2002).

¹⁷ *Supra* note 14, at 421–422.

In addition to matters of validity, automated scoring raises questions about public acceptance. Researchers Mark Shermis, Chantal Mees Koch, Ellis Page, Timothy Keith, and Susanmarie Harrington note at the end of their study: “Because the acceptability of computer ratings will be a long-term issue (as it has been with other technology innovations, e.g., computerized adaptive tests), human raters will most likely be used as a second rater to protect against bad-faith essays, cheating, and so on. Eventually, we foresee a time when the computer will be used as the only grader for low-stakes activities; but again, this will take some time.”¹⁸ (Certainly, the potential for cheating is a legitimate concern; in one statewide scoring with which I was involved, two students had, in the apparent belief their papers might be scored by a computer, done nothing more than copy over repeatedly the writing assignment.)

In fact, Anne Herrington and Charles Moran explore the profound question of what impact computer scoring will ultimately have on students’ writing when the audience for that writing is a computer—when students are, in effect, not just writing on a machine but to a machine; they conclude from their own experiment that writing to a machine desensitizes writers.¹⁹ As I myself have argued in a previous work, it is critical that the interchange between writer and reader so fundamental to written communication not be devalued by the replacement of human scorers with machines.²⁰

Uses of Holistic Scoring

The growth—and valuable impact—of holistic scoring is noted by Edward White:

To the atomization of education [holistic scoring] brought a sense of connection, unity, wholeness; to the bureaucratic

machinery of fill-in-the-bubble testing, it brought human writers and human readers; to a true-false world of memorized answers to simplified questions, it brought the possibility of complexity; to socially biased correctness, it brought critical thinking. On behalf of students, it had the human decency to ask them what they thought as well as what they had memorized; on behalf of teachers, it asked them to make complex community judgments as well as to give grades.²¹

Holistic scoring has a wide number of uses, such as in classrooms, pre- and post-evaluations of writing programs, college placement essays, statewide essay exams, and national and professional tests. It seems especially suitable for large-scale assessments in that qualified, trained readers can evaluate substantial numbers of essays quickly, efficiently, and quite reliably.

© 2004 Willa Wolcott

“To the atomization of education [holistic scoring] brought a sense of connection, unity, wholeness ...”

¹⁸ Mark Shermis, Chantal Mees Koch, Ellis Page, Timothy Keith & Susanmarie Harrington, *Trait Ratings for Automated Essay Grading*, 62(1) *Educational and Psychological Measurement* 5, 16–17 (February 2002).

¹⁹ Anne Herrington and Charles Moran, *What Happens When Machines Read Our Students’ Writing?*, 63(4) *College English*, 480–99 (March 2001).

²⁰ Willa Wolcott with Sue Legg, *An Overview of Writing Assessment: Theory, Research, and Practice* (1998).

²¹ Edward White, *Holistic Scoring: Past Triumphs, Future Challenges*, in *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations* 79, 88 (Michael Williamson and Brian Huot, eds. 1993).