

Classroom Practice Report: Holistic Scoring/Rubrics

Tony Morris

Theory:

Holistic scoring developed in the 1960s as a means of generating objective, standardized ways of evaluating student writing. Since good writing has many components, and graders have their own biases, holistic scoring requires the development of a standard rubric, or set of criteria, that takes into consideration all aspects of good writing in assigning a grade. In theory, the standardization will prevent graders from assigning too high of a grade to students with great mechanical skills in writing with very little content, for example. Likewise a rubric will not over-penalize a student with bad grammar who has a clearly organized essay.

Problems with the Theory:

There are some problems with this practice. First, there is no universal rubric; each rubric is developed by the individual teacher or institution, so what appears as a standardized set of rules may vary from one class or institution to another. This can cause anxiety in students who were taught the “rules” of one rubric, but graduate and are being held to the standards of a new one. Secondly, rubrics require a high level of training in their usage by the grader in order to apply what the rubric says to the essays being evaluated. Thirdly, rubrics are either so detailed that very few students are capable of writing at a high level, or they are so loose that the grader has difficulty differentiating from one score to the next. Additionally, students who think outside of the normal discourse are at a disadvantage because the rubric may be too restrictive to accommodate their thought processes.

Practice:

Holistic scoring is useful for graders who are grading a large number of essays. So many variables can go into how a grade is assigned. If the reader just finished a poorly written paper, the next one may seem exceptionally good even if it was just moderate. Likewise, the first paper read could be held to a higher standard than later papers because the grader recognizes the overall performance of a class or group. Having carefully developed criteria to determine grades is therefore useful.

Developing a good rubric is not an easy task. Institutions such as Harcourt and Brace spend a great deal of money over the course of years on carefully developing questions and rubrics. Those questions and rubrics are then tested to see if the students can perform to the standards of the rubrics. At times, it is determined that the rubrics were either too lenient, or too harsh, in which case the rubric is adjusted or the question is discarded. The time and resources available to Harcourt and Brace are very different than those of a high school or college educator. An experienced educator would therefore be better equipped to produce a valid rubric than a novice. However, the novice will tend to favor the students. A study by Carol O. Sweedler-Brown demonstrated that the more experienced graders tended to be more consistent in assigning grades, but that those grades tended to

be lower than the inconsistent novice grader. Her argument is that the experienced grader is more confident in their criticism and therefore more confident in assigning a lower grade.

History:

Despite the potential problems and valid fears of holistic scoring, it has become fairly standard practice. It is frequently used in placement exams, proficiency tests, and recently became a required component of the GRE. Educators as a result have begun teaching the exam. For example, educator Anne Shaughnessy has her students evaluate each other's work with the help of the Florida Writing Assessment rubric. While she argues that this practice enhances the students' classroom experience, the argument could be made that the students are hindered from practicing other genres of writing.

The literature on holistic grading has argued the pros and cons since its development in the late 1960s. Reliability quickly became the means of measuring the success of a rubric. The reliability is the ability for more than one grader to evaluate the same paper, and produce the same grade. Numerous studies have been performed in which more than one reader evaluates the same essay without knowing what grade the other(s) assigned. The objectivity of statistics has made them useful tools used to argue both sides of the argument. If a rubric has a high agreement rating, then the conclusion is drawn that it must be valid, and vice versa. However, in 1990 scholarship began to question this practice. Scholar Brian Huot reverses the binary of reliability versus validity of the practice. He notes that the discourse on how accurate the scores are neglects to consider how useful holistic scores are, and believes that the study of holistic scoring is now beginning.

Sample Rubric:

The following sample rubric is found in Clark, Irene L. *Concepts in Composition: Theory and Practice in the Teaching of Writing* (Mahwah, NJ: Lawrence Erlbaum Associates, 2003) 216. This sample demonstrates that rubrics are generally less objective in practice than in theory. For example, assigning a grade of 5 which is not as strong as a 6 assumes that the grader understands what a 6 paper would look like. Additionally, from level 6 to level 2, all of the papers could be error-free, but the higher scores do not indicate whether or not they allow for errors. This is where training comes into the practice of the rubric. If the grader has assigned a good example of each of these scores, and uses them as a reference when grading essays, it becomes more clear what grades should be assigned to which essays.

6 Range:

AN EXCELLENT PAPER: It has substantial content and clear organization and focus. It presents ideas clearly and even gracefully.

5 Range:

A VERY GOOD PAPER: The strengths outweigh its weaknesses. It has solid development and is clearly organized and focused, but it is not as strong as a 6.

4 Range:

A GOOD PAPER: The strengths of the folder outweigh the weaknesses, but the development of ideas is not as complete, the organization and focus are not as clear, and the language is not as strong.

3 Range:

A FAIR PAPER: The strengths and weakness are about equally balanced. The writer has tried to develop ideas, focus the paper, and use effective language. But parts are underdeveloped, disorganized, or confusing. The writing may also be too general or predictable.

2 Range:

A WEAK PAPER: The weaknesses outweigh the strengths. The folder is weak, underdeveloped, poorly focused, and too general. However, it could be error-free.

1 Range:

A POOR PAPER: Its weaknesses outweigh its strengths in most ways. It is unfocused, underdeveloped, and also plagued with grammatical errors that make it unintelligible.

Annotated Bibliography:

Burch, C. Beth. "Creating a Two-Tiered Portfolio Rubric," *The English Journal*, vol. 86, no. 1 (January 1997).

Burch describes her practice of not only evaluating a student's writing portfolio, but taking into account their own analysis of the work in the portfolio. The first tier is the written content, and the second tier is the student's evaluation of it. It thereby reduces the negative impact of a student turning in work that is not their best in the portfolio.

Huot, Brian. "Reliability, Validity, and Holistic Scoring: What We Know and What We Need to Know," *College Composition and Communication*, vol. 41, no. 2 (May 1990) 201-213.

Huot questions the practice of using statistics to determine the reliability of holistic scoring as the only measure of their success. He would rather see a study that questions the validity of the practice instead. Traditionally, reliability has equaled validity, but Huot finds this argument weak.

Shaughnessy, Anne. "Teaching to the Test: Sometimes a Good Practice," *The English Journal*, vol. 83, no. 4 (April 1994) 54-56.

Shaughnessy explains in her essay how she teaches her students the Florida Writing Assessment rubric before they take their proficiency exams. The students performed better on the exam, and she feels that the experience of evaluating each other's writing in class has helped the students to become better writers overall.

Sweedler-Brown, Carol O. "The Influence of Training and Experience on Holistic Essay

Evaluation," *The English Journal*, vol. 74, no. 5 (September 1985) 49-55. Sweedler-Brown performed a study in which a group of experienced educators who were untrained in the use of a particular rubric were asked to evaluate a large group of papers. Each paper was scored twice, and only if there was inconsistency of more than one grade did it go to a trainer. The trainers were well trained in the use of the particular rubric. The trainers tended to be more harsh and more consistent while the readers tended to be more lenient but less consistent.