

## Quality Assurance in Essay Scoring

Applying a score to a student essay is much different than scoring a math test or multiple choice vocabulary quiz. For the latter, the answers are black and white, right or wrong. For things like essays—which are often referred to as “open-ended” tasks—the scores are more subjective and rely heavily on human judgment. There has been an increased amount of attention paid to the issue of scoring open-ended tasks. This is due, in part, to the fact that the College Board and ACT, Inc. recently decided to include a written essay as part of the SAT and ACT exams. The stakes for students are now much higher. Additionally, there have been concerns regarding the quality, reliability, and validity of test items that are scored primarily by human judgment. Again, as the significance of test results grows, so too does the scrutiny surrounding how those results are reached. The purpose of this paper is to explain the steps taken to help ensure valid and reliable scores, and illustrate the quality assurance process for open-ended scoring.

### Based on Tradition

Traditionally, clinical exams were one-on-one evaluations that followed standard procedures but relied on the professional to interpret, administer, and score the tests. The goal of clinical human scoring is still to determine a basic, standardized structure for scoring that will ultimately lead to a common set of assessment rules for scorers to follow. Then, within those rules, there is flexibility to incorporate professional judgment and tailor

questioning to the needs of each individual student.

Human judgment in the scoring of high-stakes testing has been used for a number of years. Important clinical examinations that often lead to student interventions, as well as IQ tests, are administered, scored, and recorded by school psychologists or other professionals. In fact, in the past, those in charge of scoring clinical examinations had less training and used less structure than what is deemed acceptable today for open-ended achievement tests like student essays.

### Human Scoring of Achievement Tasks

While many of the steps related to the technical quality of assessments containing open-ended questions, such as essays, are in common with clinical assessments mentioned in the last section, there are also major differences. Scorers of open-ended achievement tests are not psychologists nor trained assessment experts. Additionally, while most clinical assessments result from one-to-one interaction between student and examiner, the scorers of large-scale open-ended achievement tests—such as essay questions on the SAT or ACT—never even see the students, much less interact with them. However, unlike clinical assessments, large-scale achievement assessments routinely incorporate additional independent scoring into the process. This step is taken to correct for scorer inconsistencies. This is why large-scale open-ended scoring is potentially more reliable and more valid than their clinical counterparts.

The following steps represent the generic “super-set” of human scoring activities in large-scale achievement tests. Some specific programs have additional steps, while some omit a few. In general, however, these steps represent the current process for large-scale assessment scoring—a process implemented and refined by PEM.

1. Scorer Selection. For large-scale assessments, all scorers have at least a four-year degree, in many cases are teachers or former teachers, and usually have experience scoring for PEM and thus are seasoned experts with our scoring and our systems. They all report to PEM professional scoring directors, who bring many years of scoring experience.
2. Scoring Rubric Generation. Scoring rubrics are the set of rules used to determine what kind of score a typical response to a question should earn. These rules are usually generated at the same time as the test is being created. Scorers use these rubrics, in conjunction with other materials and training, to score student responses.
3. Rangefinding and Anchor Set Generation. Typically, when a new open-ended item or essay question is written, it is initially tested with students. The responses to this initial test are reviewed by content experts and customer representatives, who then derive the meaning of various score points. This process is known as “rangefinding.” Papers selected in rangefinding are used to construct anchor sets—exemplar student responses that illustrate the score points on a scale as determined by the scoring rubric—and other training sets.
4. Scorer Training. Everyone involved in open-ended scoring undergoes extensive training. Scoring supervisors oversee scorers, and in most cases are themselves previous scorers who are very familiar with the PEM scoring process. Scoring supervisors are trained using the rubric, anchor set, and other training papers, and must pass a qualifying exam by identifying known, but masked “true scores” contained in student responses and identified ahead of time. Qualified supervisors then support, and often lead the training of scorers, which is a similar process supervisors are put through.
5. Backreading. Backreading is a process used to help ensure the accuracy of test scoring. Supervisors read a certain percentage of the papers that have already been scored, focusing on the work of scorers struggling with the scoring process. Supervisors can then intervene, re-train, or even dismiss scorers they find failing to score accurately and consistently.
6. Reliability (Inter-rater Agreement) Check. In many instances, student responses receive a second score. The purpose of this is to achieve “inter-rater agreement,” which simply means that the second score falls within close range to the first score. For example, some clients require 70 to 90 percent perfect agreement. These multiple readings are managed at PEM by our computerized system, and the data is analyzed as evidence of our scoring reliability. This data can

also be used by scoring supervisors to help ensure the quality of individual scorers.

7. **Resolution.** Each PEM client establishes its own target agreement rates. During the reliability check explained above, if those target agreement rates are not met, the electronic system will alert and notify the supervisor that a discrepancy exists and must be resolved through an adjudication process. Sometimes this process involves routing the paper to other scorers as opposed to going directly to a supervisor. Either way, client specifications dictate the acceptable outcome. In some cases, an average of the three scores is used. Other times, a replacement of the discrepant score is used. And in certain instances, the use of the supervisor score is used.
8. **Validity Check.** The use of “validity papers” is incorporated into the scoring process to help ensure that scoring is of the highest possible quality and reliability. These papers are randomly delivered to scorers and are indistinguishable from the actual responses of students. The purpose is to determine if all scorers are scoring according to the standards established at the rangefinding stage. Scorers must score the validity papers at an accuracy rate agreed to by PEM and the client. Scorers who fail to do so may be re-trained and/or removed from the scoring process.

## Summary and Conclusion

Scoring student essays and other open-ended tests with consistency and reliability is, indeed, a complicated task. The prevalence of these kinds of achievement tests continues to grow—ensuring the highest technical quality of scoring is paramount. Some people are surprised by the depth of the quality assurance processes used in the scoring of essays and other open-ended tests. These processes are similar to those that historically have been used for clinical assessments. However, in many ways they are superior to their clinical counterparts because of the additional scorings and resolution of discrepant scores. At PEM we have an automated quality assurance process that quantifies and then systematically reduces the error interjected by human judges.

- Jon S. Twing, Ph. D.
- Daisy Vickers

Pearson Educational Measurement is the largest comprehensive provider of educational assessment products, services and solutions. As a pioneer in educational measurement, PEM has been a trusted partner in district, state and national assessments for more than 50 years. PEM helps educators and parents use testing and assessment to promote learning and academic achievement.

Measurement operates as a business of Pearson Education, the world's largest education company. More information is available at [www.PearsonEdMeasurement.com](http://www.PearsonEdMeasurement.com).