MEASUREMENT INCORPORATED

EXCELLENCE IN ASSESSMENT™

The Quest for Consistency: Double-Scoring Policies and Impacts on Fairness

Corey Palermo, Ph.D.
Measurement Incorporated

Background

Large scale summative assessments frequently include constructed response items, requiring students to produce written answers. Unlike multiple choice and other "objective" selected-response item types, constructed responses allow students to provide an endless array of possible answers. Consequently, the quality of students' constructed responses is evaluated using scoring rubrics and exemplar responses. This evaluation can be performed by trained human raters and/or automated scoring systems. Rigorous rater training and the establishment of clear, detailed scoring rubrics are foundational steps to minimize scoring discrepancies.

A common practice across state assessment programs is to double-score a portion of constructed responses, typically 10–20%. This double-scoring allows for reporting inter-rater agreement, which measures the consistency of ratings between two or more evaluators, whether human, machine, or both.

Upon encountering discrepancies in scores—for instance, if one rater assigns a 0 and another a 1 to the same response—the natural inclination may be to rectify these differences, which indicate a lack of consistency. It may seem particularly important to intervene in cases where the scores are not adjacent (e.g., one rater assigns a 0 and another a 2 to the same response).

The Problem

In the common case where only a proportion of responses are double-scored (i.e., second read), resolving score discrepancies introduces a risk to fair and valid interpretations of test scores. To understand the implications, it is essential to consider the concept of fairness as outlined in the Standards for Educational and Psychological Testing (the *Standards*, AERA, APA, & NCME, 2014).

According to the *Standards*, fairness is a critical aspect of validity, crucial for ensuring that inferences drawn from test scores are comparable across test takers. The *Standards* emphasize that "test takers should receive comparable treatment during the test administration and scoring process" (p. 65). The cornerstone of score validity is uniform and consistent procedures.

In the context of the *Standards*, it becomes evident that resolution methods—such as involving a third rater or having a scoring supervisor resolve non-adjacent scores—compromise fairness and score

comparability when not all responses are subject to a second review. Specifically, this practice advantages a fraction of students who benefit from a more robust scoring process than others.

An Example

Consider two students, Casey and Leila, who each submit a similar-quality response to a question. Casey's response is among 20% which are double-scored and receives initial ratings of 0 and 1 from two raters. The discrepancy triggers a resolution process, and Casey's final score is adjusted to 1 following a supervisor review. Leila's response, not selected for double-scoring, is reviewed by a single rater and receives a score of 0.

This scenario illustrates inequitable treatment during the scoring process: Casey benefits from the chance of a score adjustment through the double-scoring and resolution process, while Leila, who provided a response of similar quality, suffers as a result of not receiving the same opportunity.

Recommendations

Adopting a uniform protocol for handling scoring discrepancies will help ensure that each response is treated equally, thereby supporting fairness and validity. If a goal of the program is to maximize scoring accuracy, this can be achieved while maintaining score comparability by double scoring and resolving all responses. Alternatively, statistical methods can be used to adjust for any scoring inconsistencies found. For example, Item Response Theory (IRT) can be applied to adjust scores to a common scale that accounts for rater discrepancies, thus maintaining score comparability. Ultimately, scoring policies in state assessment programs will strike a balance between ideal practices and real-world constraints.

Importantly, many strategies are available to ensure score quality without compromising fairness. Foremost among these is rigorous evaluation of rater accuracy. This is best achieved using validity responses, or pre-scored benchmark responses distributed to raters amongst operational responses. Because the benchmark scores are determined by experts, validity responses offer an external reference for evaluating rater accuracy by measuring the alignment of raters' scores with the benchmark scores. Additional strategies include systematically rescoring responses most likely to be scored inaccurately. This includes responses scored by raters with poor validity performance. Rather than being resolved, double-scored responses with scores that clearly indicate error, such non-adjacent scores or a mismatched scorable/nonscorable (for example, off topic and valid score) can be reset and rescored in the same way. These steps can be used to address rater drift and ensure score quality while maintaining test score comparability.

Conclusion

Double-scoring a proportion of responses serves the goal of measuring scoring consistency, but when the process includes third readings or resolutions score comparability is undermined. Alternatives that ensure scoring quality and fairness should be a priority in assessment programs, aligning with the foundational principles of fairness, validity, and appropriate use of scores as outlined in the *Standards.*