Halo Effects and Analytic Scoring Research Report Summary

Emily R. Lai
Edward W. Wolfe
Daisy H. Vickers

As we move into the next generation of work, it is clear that performance assessments will be an important component of what is to come. Due to ongoing concerns about the reliability of human scores and the cost of obtaining those scores, stakeholders will desire the most information possible from student responses to assessment tasks.

Examining Scorer Bias

This study focused on two related topics: the potential for bias in scores assigned to student writing due to scoring process and how much useful information can be gained by assigning multiple scores to student essays. The potential bias in this context is the halo effect--a scorer's failure to differentiate distinct features or characteristics of an essay when multiple analytic scores are assigned to that essay by a single rater. For instance, if an essay has excellent idea development, a reader might be biased to give it higher mechanical scores than it deserves.

Study Methodology

The study compared scores from four groups of scorers, each assigning a single analytic score for a specific skill being assessed within student writing responses (e.g., conventions, organization, development, or voice). In a fifth group, each scorer assigned all four analytic scores to a particular student essay. Finally, a sixth group of scorers assigned a holistic score to student writing responses using a rubric that merged all four writing features into a single score.

Analysis of Halo Effect

The results suggest that human scorers exhibit halo effects when assigning multiple analytical scores to student writing responses. Specifically, scores of the single analytic scoring groups were less similar between traits than were the scores assigned by the group that scored all four traits at one time. The study results also suggest that only two scoring factors, idea development and mechanics, are distinguishable in these middle school expository essays.

Scoring Implications

The results are important to consider as the assessment industry evolves and includes more constructed-response items and performance-based assessments, particularly ones that focus on multiple-dimensions of student performance. Although a single-trait-per-rater scoring design may minimize the risk of the halo effect, that design significantly increases the costs associated with scoring, so additional research is needed to determine whether rater training and monitoring can also reduce the halo effect. In addition, test developers should carefully consider how to best depict the traits for which scores are assigned to writing assessments and other performance assessments because our results raise compelling questions about the number of traits that raters can differentiate. This is especially important in light of the fact that assigning fewer scores to a student response would lower scoring costs.

Our results raise compelling questions about the number of traits that raters can differentiate.