Reliability refers to the consistency and dependability of an assessment. A reliable assessment provides a reasonably consistent picture of what students know, understand, and are able to do regardless of when or where students were assessed.

Several factors can affect a performance assessment's reliability, including its administration, the clarity of the performance task, and its evaluation or scoring.

Clarity of Task

Often a performance task may include vague language, especially when the performance outcomes are not explicitly stated. Clearly written tasks focus the respondents on the expected performance and generate responses that can be scored consistently. To improve the reliability of the student responses, when writing the task, consider what you are looking for as evidence of achievement. Refer to the performance outcomes and examine the content and cognitive behavior. Align the directions with the performance outcomes and indicators and include the scoring rubric so that students know the expected levels of performance.

EXAMPLE

A task requires students to explain the Boston Tea Party. While observing students working on their essays, the teacher notices that some students are simply listing major facts while others are detailing the cause-effect relationship among the major events. The teacher stops students and revises the task to "Identify key events that led to the Boston Tea Party. Which groups were most affected by it and how? Justify your answer." She has revised the flawed assessment during administration to avoid receiving responses that were not a fair measure of students' knowledge and skills due to the vague directions. (Example adapted from Chatterji, 2003)

Administration

To maximize reliability of a performance assessment, provide clear student instructions as well as teacher instructions for other teachers who may administer the assessment. When possible, provide practice activities or tasks to support students' understanding of what will be required during the performance assessment demonstration. These steps will help ensure that the assessment is a reliable measure of student performance.

Evaluation

Other aspects of an assessment's reliability are how the assessment is evaluated, scored or rated and the extent to which the ratings are the same across different raters. Student responses should be evaluated consistently, using the same criteria, so that appropriate inferences may be made about students' knowledge and skills. Evaluators can improve the consistency in which they apply the same criteria across responses by using well-described rubrics, anchor papers consisting of sample student work representing different performance levels, and by participating in calibration sessions. Calibration sessions are meetings during which scorers are trained on using the rubrics, apply the rubrics to example responses, and discuss their thought processes until they reach consistency in their scoring.

EXAMPLE

Over several months a team of elementary teachers used their common planning time to develop a science performance assessment. Students must carry out a scientific investigation and summarize their findings. One of the tasks is an open-ended response on the importance of inquiry in science. One of the teachers piloted the performance assessment with her class and collected the student responses. The group of teachers uses these responses and the assessment's rubric to calibrate themselves—meaning they independently provide the same rating to the same student responses.
This level of consensus is achieved through a formal calibration process that entails reviewing several samples of student work, rating the work, and discussing the ratings until an agreement in scores is reached. For more information on calibrating student work, see the calibration protocol from the Rhode Island Department of Education (2013).

Questions for Consideration

(1) Is the performance assessment task clearly stated?

(2) Does the assessment yield reliable results every time you administer it?

(3) Is the assessment reliable for ALL students, including for each subgroup?

(4) Do different raters arrive at the same rating for the same student response?

References for Reliability

Chatterji, M. (2003). *Designing and using tools for educational assessment*. Boston, MA: Allyn & Bacon/Pearson.

Klinger, D. A., McDivitt, P. R., Howard, B. B., Munoz, M. A., Rogers, W. T., & Wylie, E. C. (2015). *The classroom assessment standards for preK-12 teachers.* Kindle Direct Press.

Stiggins, R. (2007). Assessment through the student's eyes. *Educational Leadership, 64*(8), 22–26.

Rhode Island Department of Education & The National Center for the Improvement of Educational Assessment. (2013). *Calibration protocol for scoring student work. A part of the assessment toolkit.* Providence, RI: Author. Retrieved from http://www.ride.ri.gov/Portals/0/Uploads/Documents/Teachers-and-Administrators-Excellent-Educators/Educator-Evaluation/Online-Modules/Calibration_Protocol_for_Scoring_Student_Work.pdf