

What About 1-2?

Tests and quizzes are the only way to make sure that our assessment is objective and fair. All this other stuff is too subjective.

Tests and exams are only one way to see what a student knows. And not everything that is expected of students in the curriculum can be measured by a test. For example, students cannot be evaluated by a paper-and-pencil test in curriculum areas where they are expected to communicate orally.

Tests are not objective measures. They usually result in scores that represent a quantitative measuring system related to numbers, which is as close as a test comes to being objective. Think about this: Two physics teachers might be preparing a test for the same unit of study. One teacher creates a test of 12 questions that appeal to her and the other teacher

builds a test of five different questions that have served him well in the past. By selecting questions that they view to be important, these two teachers have built in subjectivity.

Instead of asking ourselves whether our measures are objective, we need to be asking ourselves whether our measures are reliable and valid. Reliability refers to repeatability. Can the student show what he knows in different situations and at different times? Validity refers to the match of the evidence of learning to what is to be assessed – what is to be learned. And, to illustrate, let us go back to the example of an oral presentation. We cannot evaluate whether a student can orally communicate ideas to an audience by asking them to complete a paper-and-pencil test. From a classroom assessment perspective, this evidence of learning is not valid, given what was to be learned.

Teacher professional judgment is more reliable and valid than external tests when teachers have been involved in examining student work, co-constructing criteria, scoring the work, and checking for inter-rater reliability (ARG, 2006, Burger et al., 2009).

Inter-rater reliability is defined as the result of learning to make an informed professional judgment. Educators engage in a process of inter-rater reliability when they meet, create quality criteria, and build a scoring rubric for student work. The student work could include, for example, a performance task, a product, observations of application, or a body of evidence. Each educator examines and then scores the student work using a scoring rubric.

At this point, all the scores are examined for consistency among all the educators. The percentage to which they agree is used to determine inter-rater reliability. The higher the percentage of agreement among all the educators' rating (i.e., the more the scores are similar), the higher the inter-rater reliability will be. This process helps teachers refine and improve their professional judgment.