Reliability for Teachers Activity: What are some of the issues related to reliability?

This activity will help you answer the essential question:

- **What are some of the issues related to reliability?**

Activity 2: What are some of the issues related to reliability?
You may complete this activity individually or in groups

Read the following excerpt from *A Primer on Setting Cut Scores on Tests of Educational Achievement* and then complete the following activity:

Identify the reliability issues identified. Discuss or reflect on the following: What are the implications for understanding the meaning of cut scores? What are the implications for you, your students, and your school?

Reliability of the Classifications Made on the Basis of Cut Scores

From: *A Primer on Setting Cut Scores on Tests of Educational Achievement* by Michael Zieky & Marianne Perie, ETS, 2006. Available at: http://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf. Retrieved March 7, 2012.

A useful synonym for "reliability" in an assessment context is "consistency." If the student stays the same, how consistent will her scores be if she takes an alternate form of the test (different questions covering the same content at the same difficulty)? For tests that are used with cut scores, it is important to get answers to the following questions:

- What proportion of students would be classified the same way if they had taken a different form of the same test?
- What proportion of students would be classified the same way if they had taken the same form on a different day (assuming no changes in knowledge)?
- What proportion of students would be classified the same way if their responses to the constructed-response questions, such as essays, had been scored by different people?

All of those questions deal with the reliability of the classifications being made. The reliability of classification will not be perfect, even for good tests. Every test is only a sample of all the questions that could be asked. Test takers are not likely to be equally knowledgeable about all of the legitimate questions that could be asked, so test form to test form differences are likely. Day-to-day fluctuations in students' attention, memory, luck in guessing and so forth are also expected.

Even scorers who have been well trained will disagree occasionally about papers that are near the borderlines of score differences. For example, a response that one judge considers a high 2, another judge may legitimately see as a low 3. This discrepancy becomes a problem if the cut score requires a minimum of 3 to be classified as proficient. The discrepancy would not affect the reliability of classification, however, if scores of both 2 and 3 were considered basic.

Students with similar scores on a test tend to be similar in what they know about the tested subject. Most tests cannot distinguish well between students with scores that are very close to one another. Whenever a cut score is used, however, students with scores just above the cut score and students with scores just below the cut score will be classified differently. What this means is that students who score near the cut score may pass or fail a test because of random fluctuations.

The more reliable a test is, however, the less likely it is that the scores will be affected by large random fluctuations. All other things being equal, longer tests will be more reliable than shorter tests; and objectively scored tests will be more reliable than subjectively scored tests. Note that if the reliability of classification is such that test takers are classified the same way 50% of the time, the same level of consistency could be achieved by flipping a coin. At that level of reliability, the test is providing no useful information about the proficiency levels of individual students.