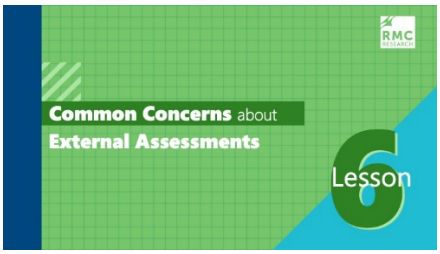

Time: 14 minutes 12 seconds

Index

Lesson Overview

In this lesson you'll be introduced to:

- the utility of external assessments;
- common concerns about external assessments and how to respond to them in your own practice;
- the limitations of performance level reporting;
- recommended activities and suggested resources (which are expanded on in this facilitator guide).

Video Outline

- Introduction (00:00-01:51)
- Utility of External Assessments (01:52-07:30)
- Limitations and Misuse of Performance Level Reporting (07:31-12:25)
- Putting It All Together (12:26-13:08)
- Activities and Resources (13:09-14:13)
    - Recommended Activities
    - Suggested Resources

Your Learning Objectives
Record your objectives and points to focus on.

Your Stops
Make notes on stopping points and content discussion you would like the participants to take part in.

Stopping Point	Content Discussion	Notes

Suggested Discussion Questions

- ESSA, like NCLB, requires states to use multiple measures to cover higher order thinking skills, and further requires the states to produce "individual student interpretive, descriptive, and diagnostic … as soon as practicable." How well do you think your state has complied with these requirements? Explain.
- Several years ago, officials in a group of southeastern states questioned why their National Assessment of Educational Progress results in reading and mathematics (percentages of students scoring proficient or above) were so much lower than the results on their own state tests. Evaluate each of the following three explanations:
    - The state tests were better aligned with curriculum and instruction in those states.
    - The state tests were easier than the national tests.
    - The state standard setting procedures led to threshold scores (or cut scores) that produced the differences in results that were obtained.

Recommended Activities

- For each of the following issues, prepare a list of brief, bulleted "talking points" to assist you in responding to colleagues, parents, and others who raise the issues.
    - These tests don't cover what I teach.

- The tests are too long and tell us too little.
- We're testing students too much!
- It takes so long for us to get test results that they are no longer useful.
- Why do we get conflicting results from different testing programs?

Suggested Resources

Kahl, S. (2019). *Expectations of State Assessments: More Information from Shorter Tests*. Posted paper, Portsmouth, NH: RMC Research Corp. https://www.assessmentworkshop.com/wp-content/uploads/2021/09/ExpectationsV2.pdf

Kahl, S. (2015). *Proficient, Eligible to Graduate, College Ready? The Mystery of Achievement-Level Assessment Results*. Posted paper, Portsmouth, NH: RMC Research Corp. https://www.assessmentworkshop.com/wp-content/uploads/2021/09/The-Mystery-of-Achievement-Level-Assessment-Results.pdf

Shepard, L.A., Davidson, K.L. and Bowman, R. (2011). *How Middle School Mathematics Teachers Use Interim and Benchmark Assessment Data*, CRESST Report 807. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing at UCLA. http://cresst.org/wp-content/uploads/R807.pdf

Layton, L. (2015). *Study says standardized testing is overwhelming nation's public schools*. Washington, DC: Washington Post. https://www.washingtonpost.com/local/education/study-says-standardized-testing-is-overwhelming-nations-public-schools/2015/10/24/8a22092c-79ae-11e5-a958-d889faf561dc_story.html

Script

Slide 1

<blank>

Slide 2

- Recall from Lesson 1 that external assessments are from sources external to the classroom such as the state or a commercial test publisher.
- They include
    - interim general achievement measures, which cover the full range of grade-level material in a subject,
    - interim benchmark tests, which focus on just a portion of a content domain, and
    - end-of-year summative tests, which, like interim general achievement measures, draw from the full range of grade-level content.
- These external tests are typically designed for the identification of students or curricular areas needing additional attention, program evaluation, and school accountability.

Slide 3

- This lesson deals with issues, such as those shown here, associated with “external” assessments.
- At the conclusion of this lesson you will
    - have a better understanding of the different sides of these issues,
    - be able to respond accurately to others who raise them, and
    - have a greater appreciation of the important roles these assessments play in a balanced assessment system.

Slide 4

- The information in this lesson will help you
    - interpret standardized test data correctly,
    - best utilize external testing results, and
    - respond to common concerns about testing raised by parents, colleagues, and others.

Slide 5

- This lesson is organized into two main sections. The first addressing concerns about the utility of external assessment results. The second dealing with limitations of performance level results.
- The final section will introduce suggested activities and accompanying resources.

Slide 6

<blank>

Slide 7

- This is a common concern raised by teachers.
- It’s important to point out that instruction and assessment should address the same curriculum standards. In the case of state assessments, a great deal of effort is made to assure that the tests align with state curriculum standards.
- However, it is true that if educators focus on higher order thinking skills, then there might be a partial disconnect. “Efficient” state tests are typically dominated by item types that do not measure these skills well.
- Unless there is some flexibility in the scheduling of testing, some external benchmark tests could be out of sync with local instructional sequences.

Slide 8

- If there’s a mismatch between the state’s external test and what you teach, there would likely be a need for curriculum review. During that activity, local educator teams should consult the curriculum standards published by the state department of education.

- If there is a mismatch between external tests and what you teach because you give more attention to higher order thinking skills, then you should be commended. If you do not yet give much attention to higher order thinking, then reference Lesson 3 where we provide some ideas that can help you assess higher order thinking skills.
- Your instruction and district-wide benchmark testing should be coordinated. If instruction and state benchmark testing is out of sync, take that into account when interpreting results.

Slide 9

- A common complaint from educators and parents is that standardized tests, especially state tests, are too long.
- Yet educators often want more detailed information from these tests to help them make the best instructional decisions.
- In short, they want more information from shorter tests.

Slide 10

- Recall from Lesson 3 that test reliability and validity are closely tied to how well the test items and tasks sample the target content and skills.
- Standardized tests are as short as they can be, while still producing a reliable total test score.
- An individual student's subtest score, based on a subset of items in the test, is not particularly reliable. A school's average of students' subtest scores is more reliable and therefore more useful for program evaluation.
- In other words, teachers should not be looking to most standardized tests for individual student diagnostic information.

Slide 11

- Research has shown that overuse of external interim tests is a major contributor to the much publicized over testing problem.
- If there is little instructional time between administrations of general achievement measures then there is little academic growth. Measurement error can overshadow growth at the individual student level and negative growth could be falsely attributed to some students.
- Interim benchmark tests, given perhaps three times per year and addressing just a portion of the year's instructional content, occur "after the learning," and are therefore like summative tests.
- Teacher's own summative tests at the end of units or marking periods are likely better aligned with instruction.
- Thus, external benchmark testing may be duplicative or unnecessary. However, it could be used for identifying remedial needs or for program evaluation.

Slide 12

- First, make sure you are using external test results for the purposes they were designed for.
- A worthwhile summertime activity could be for teams of teachers, with the support of administrators, to review all the testing that takes place in the district. Teachers should ask

themselves if the results from a particular external test provide greater value than classroom summative tests and other external tests used.

Slide 13

- It should be clear by now that most external tests, whether commercial or state, are not designed to inform teachers' day-to-day instruction. For program evaluation purposes, immediate results are not so critical. In fact, it is suggested that patterns of results across years should be used to make important programmatic decisions. **This is particularly true for small schools whose results at a particular grade level are often unstable. They reflect the varying abilities of the students passing through a tested grade in different years.**
- Tests that produce the quickest results are tests that rely on immediate machine-scoring. These tests generally focus on low level cognitive skills. Deeper learning – the ability to apply such skills to more complex tasks – should be an important goal of instruction and should also be tested.

Slide 14

- Generally, external tests are not designed to produce results that guide day-to-day instruction.
- Develop effective formative assessment practices. These are what lead to the diagnosis of individual student learning gaps and adjustments to instruction *during the learning*.
- Your formative assessment and classroom summative assessments should address both low order and higher order thinking skills.

Slide 15

<blank>

Slide 16

- In Lesson 5, we explained the need for reporting both scaled scores and performance levels in standardized testing programs.
- We saw how two students could perform very differently on a test, yet be assigned to the same performance level.
- And we saw how two students could score almost the same on a test and be assigned to two different levels with different descriptions of capabilities.
- So what does this mean for the classroom teacher?
    - First, recognize that performance level reporting -- that is, the percentages of students at various levels -- is particularly useful for program evaluation.
    - But in conversations with individual students and parents, *where* the students' scores fall within a performance level is important to point out.
    - Instructional activities should not be built on the assumption that students at the same level are at the same place in their learning.
    - Next higher levels provide reasonable goals for students to work toward.

Slide 17

- Another problem occurs when individuals make inaccurate claims about testing programs, either knowingly or otherwise, based on performance level results.
- Here are two examples.

Slide 18

- Several years ago, two university professors widely publicized their concerns about a state's assessment program.
- They argued that results on the state tests in reading and math were not correlated with the results on a widely used and respected commercial test.
- Their evidence for this claim was a discrepancy in the percentages of students proficient or above reported by both programs.
- The commercial test put approximately 75 percent of students in those two upper performance levels, while the state test put only about a third of the students in those categories.
- Therefore, they concluded there was something wrong with the state tests.

Slide 19

- The scaled scores from the two programs were actually well correlated. The seeming discrepancy was primarily due to where the cut scores for the different performance levels were placed.
- Looking at this graphic, it would not be unlikely that some students scored Novice (N) on the state test and Proficient (P) on the commercial test.
- Differences in the percentages of students at various performance levels have nothing to do with test quality. Furthermore, one could not even claim that the commercial tests were easier than state tests because of the higher percentages proficient and above.
- Again, the seeming "discrepancy" was mainly due to where cut scores were established.

Slide 20

- The second example of the misuse of performance levels occurred in a state deciding whether to transition to a new testing program.
- Those favoring the new program argued that the existing program was not predictive of college readiness or success. They further argued that the old program reported high percentages in the upper categories even though college officials continually complained that high percentages of incoming freshman required remedial course work. The new program reported much lower percentages of students scoring in the upper performance levels.
- An independent research group was engaged to study the problem. Performing appropriate analyses on scaled scores, the researchers concluded that the two programs predicted college readiness and success equally as well. Again, the issue was where cut scores had been established. When the performance standards (cut scores) were first set for the exiting program, the likelihood of success in the first year of college was not a consideration for the upper performance levels.

Slide 21

- Thank goodness performance levels have gotten us away from the normal "bell" curve! …. How many times have we heard that?
- Generally, numerical scores on tests, just like many other things in our world, are normally or close-to-normally distributed.
- Without the curve, this display could be quite similar to one in a parent report from an external summative assessment, a state assessment for example.
- But the curve shows what's really happening under the surface. The scaled scores are close to normally distributed. Those scores are determined before score ranges for performance levels are created.
- The curve shows what the percentages at the bottom show – that the vast majority of student scores fall into the Basic and Proficient levels.

Slide 22

- There are two big ideas underlying the content of this lesson:
    - First, there are a lot of myths and misconceptions regarding external assessments; and with a better understanding of them, practitioners will be more likely to use different measurement tools for their intended purposes and stakeholders will better understand their results.
    - Second, for various decisions about students, lines have to be drawn that separate higher from lower performance on tests. But distinctions like pass/fail or proficient-or-not do not tell the whole story about student achievement.

Slide 23

- In anticipation of questions, it is generally useful to prepare a list of talking points in advance. This activity asks teachers to prepare a list of talking points for each of several frequently raised issues regarding testing.

Slide 24

- The Washington Post article by Layton discusses the problem of over testing in U.S. schools and the causes of it.
- The second reference is to a research report by Shepard et al. It presents interesting findings on how and how much interim general achievement and benchmark assessments are used.
- Two other resources listed in the Lesson 6 Supplement are posted papers by the primary author of the video lessons. They deal with several of the issues commonly raised about testing that are summarized in Lesson 6.