Time: 16 minutes 48 seconds

Index

Lesson Overview

In this lesson you'll be introduced to:

- how multiple-choice, constructed-response, and performance tasks work;
- how students respond to different test item types and the implications for test reliability and validity;
- when and why to use different test item types;
- suggestions for teachers to improve their test creation; and,
- recommended activities and resources (which are expanded on in this facilitator guide).

Video Outline

- Introduction (00:00-01:20)
- Introduction to Common Test Items (01:21-03:36)
- Test Item Types and Content Validity (03:37-09:49)
- Value of Using Multiple Test Item Types (09:50-14:39)
- Putting It All Together (14:40-15:14)
- Tips for Your Role: Teachers (15:15-15:51)
- Activities and Resources (15:52-16:48)
    - 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

- What are several reasons to use a variety of item types in a test or throughout a course?
- What's wrong with the statement, "Multiple-choice tests are more reliable than constructed-response tests."?

Recommended Activities

- Access the test item questionnaire.
- The questionnaire was developed to be a discussion starter. Find a colleague who will take the questionnaire at the same time you do. After the two of you complete it, discuss both the directions and each of the items in light of the content of the video lesson.

Suggested Resources

Scantron (2020). *Item Types and Cognitive Complexity Models in Higher Education*. Blog post, Eagan, MN: Scantron. https://www.scantron.com/blog/item-types-and-cognitive-complexity-models-in-higher-ed-assessment/

Shepard, L.A. (2010). What the marketplace has brought us: item-by-item teaching with little instructional insight. *Peabody Journal of Education*, London: C. Taylor & Francis Group, 85: 246-257 https://doi.org/10.1080/01619561003708445

Note: We recommend teachers acquire one of the two publications listed below, or a similar comprehensive resource for teachers, for use throughout their teaching careers.

Taylor, C. S. and Nolen, S. B. (2008). *Classroom Assessment: Supporting Teaching and Learning in Real Classrooms (2nd edition)*. Upper Saddle River, NJ: Pearson Education, Inc.

Nitko, A. J. and Brookhart, S. (2019). *Educational Assessment of Students (8th edition)*. New York, NY: Pearson Education, Inc.

Script

Slide 1

- Welcome to Lesson 4 where we take a closer look at different types of test items and how students respond to them.

Slide 2

- There are many different test item types.
- Here you see an example of multiple-choice on the top and constructed-response on the bottom.
- In this lesson you'll see that different test item types yield different kinds of information about student learning.
- You'll see when and why to use different item types, so you can get a better understanding of what your students know and can do.

Slide 3

- This lesson has three main parts.
- We'll start by digging deeper into how different test-item types work.
- We'll also describe:
    - the relationships between item types and test reliability and validity,
    - issues associated with machine and human scoring of student responses,
    - and the information item responses yield.

Slide 4

- This topic is important because it can lead to a better understanding of the importance of using a variety of test-item types
- The result will be a deeper understanding of students' capabilities and instructional needs, and ultimately better instructional decision making and higher student achievement.

Slide 5

<blank>

Slide 6

- In this lesson, we'll focus on three commonly used test-item types that are particularly helpful to teachers:
    - multiple-choice items, for which students choose a response from a series of options;
    - constructed-response questions, such as essay questions or questions requiring students to explain their answers or show their work;
    - and performance tasks, such as student oral presentations, written reports, projects, or investigations.

Slide 7

- Multiple-choice items are the most frequently used items today. Many of them can be administered in a short period of time, and they are easy to score.
- There are other types of selected-response items, such as true/false and matching. We will not be addressing them here because multiple-choice can generally measure knowledge and skills better.

Slide 8

- As we said previously, constructed-response questions require students to explain their answers or show their work.
- The constructed-response questions we refer to in this lesson often require 8 to 10 minutes to complete and a half page of space for responses. They are best scored by humans.
- The students' written responses or work displays are typically worth between 0 to 4 points, depending on their quality. (Four-point questions are often used in external assessment programs that go beyond the measurement of low-level skills. They are not extremely time-consuming to answer or score.)

Slide 9

- People use the term "performance task" to mean different things. As part of a proctored timed test they are typically more like longer constructed-response measures.

- For the purposes of this lesson, however, we'll be thinking of performance tasks as consisting of much longer activities, such as projects or experiments, for which students produce scorable products or demonstrations.
- Such assessments may be curriculum embedded – that is, integral parts of the curriculum –and are almost always best scored by humans. They would typically be worth many more than 4 points each.

Slide 10

<blank>

Slide 11

- Recall from Lesson 3 that a test with content validity includes a "wide enough sample" of items to show that students understand the full range of learning objectives within a unit or marking period.
- Tests should be aligned with or "cover" curriculum standards.
- Curriculum standards identify learning objectives, which address both high and low level skills.
- Thus, to be valid, a comprehensive test or a set of less comprehensive tests administered over time should address a full range of thinking skills.
- Concepts, topics, and low-level skills can be addressed by multiple-choice items, but a variety of item types are needed to cover both low-level and higher-order thinking skills. Higher order thinking reflects "deeper learning" – the application of knowledge and skills to deal with complex problems or situations.

Slide 12

- In Lesson 3 we described Webb's Depth of Knowledge levels.
- We noted that most multiple-choice questions test low-level thinking skills, while constructed-response and performance tasks tap higher order skills.
- Here we present the four depth of knowledge levels with the most appropriate test item types. <pause>

Slide 13

- When taking a multiple-choice test item, students will come up with many more answers than the four options the test developer offers them, even though the incorrect options should correspond to common misconceptions. Percentages of students selecting different options are not always as revealing as people think.
- When students do not find their answers in the options, they may spend more time on it, skip it, or guess. If they guess correctly (there's a one in four chance of that), they get full credit for that – one point.
- Sounds like a scoring error, doesn't it? It certainly is equivalent to a scoring error, even though there is no error in capturing the answer the student marked.

Slide 14

- Percentages of students selecting each of the options in a multiple-choice item can be useful, but over-interpreted.
- For example, if 45 percent of the students in a class picked 2/5 [two-fifths] for the answer to ½ + 1/3 [one half plus one third] we know there's a general problem and what it is.
- Option percentages have their uses, but sometimes, teachers can learn just as much, if not more, from the short answer format.
- This is particularly true if the item-writer is not proficient at developing multiple-choice items.

Slide 15

- If you score your multiple-choice tests by hand, you might consider going with the short-answer format. Short-answers are just as quick to score and you'll have eliminated guessing. You may get a better sense of the students' errors and misconceptions.
- The test scores will be lower, because students can't guess the correct answers or partially guess by first eliminating obvious wrong answer options. But the scores will be better reflections of what your students know and can do. Also, the development of good multiple-choice questions is not as simple as many think.

Slide 16

- Many multiple-choice questions are surrogate or indirect measures of the knowledge or skills of interest.
- Let's look at some examples.

Slide 17

- Think about these two test items, which present the same problem in both multiple-choice and constructed-response format. <pause>
- With multiple-choice options given, the quadratic equation can be solved using a low-level algebraic skill by substituting the answer options for x and simplifying.
- In constructed-response format, the student has to know how to solve a quadratic equation – the tougher target skill – to get the correct answer. Also, the teacher can see the students' work on the problem.
- While multiple approaches may sometimes be a good thing, that is not the case here, if the goal is really testing the ability to solve a quadratic equation. Even the students who have the target skill may still use the lower level approach to answer the multiple-choice version of the item. And you would never know because of the multiple choice item format.

Slide 18

- Now look at the reading comprehension item presented in both formats:
- There's a big difference between a student selecting the correct conclusion provided in Item Format A and coming up with the correct conclusion on his or her own in Item Format B. That would be even more the case if the constructed-response question asked the student to justify his or her conclusion with evidence from the story.

Slide 19

- It's important that the questions you ask your students are measuring what *you* want them to measure.
- Professional test developers sometimes field test items, then interview students to find out how they arrived at their answers.
- The best way for you to do that is to think hard about the thought processes your students will apply to answer a question.

Slide 20

- While multiple-choice items definitely have their place in student testing, some multiple-choice fans don't see the added value or capabilities of constructed-response items and performance tasks.
- Those who think you can do the same with multiple-choice are missing a key point. A good constructed-response question or performance task requires the application of multiple skills or pieces of information. Multiple-choice questions often isolate them. It is not true that being able to demonstrate the specific skills in isolation guarantees that a student can pull them all together to deal with a complex problem or task.

Slide 21

<blank>

Slide 22

- In this section we will see how the use of a combination of test item types is good practice.
- Recall from Lesson 3, when it comes to coverage of standards by a test, having enough items measuring something is reliability. Using enough items covering the right stuff is content validity.
- Good coverage of the topics within a content domain of interest can be achieved with a lot of multiple-choice items, but an all multiple-choice test would likely shortchange higher order thinking. Such an instrument could be reliable, but not valid if the intent was to measure both high and low-level cognitive skills.
- A more valid, and still reliable, measure could be achieved with many fewer 4-point constructed-response items, but more effort would be required to score them.
- Many people have compared a 50-item multiple-choice test to a single performance task and claimed that performance testing is not reliable. That's an unfair comparison. Yes, reliability depends largely on the number of items or tasks, and it would take more than two or three performance tasks to provide adequate coverage of a content domain for reliable and valid scores. That's why the use of a combination of test item types is often good practice.

Slide 23

- When humans score students' constructed responses, they don't always give responses the right scores. They make scoring errors. How can the constructed-response format yield the same test reliability as multiple-choice?
- The short answer to this question is that: Item for item, a 4-point constructed-response question provides much more information about the test takers than a multiple-choice question. In terms of scores and information produced, a single 4-point constructed-response question is worth four multiple-choice questions.

Slide 24

- Remember, a 4-point constructed-response question taps more knowledge and skills than your typical multiple-choice item. Furthermore, the possible scores (0, 1, 2, 3 or 4 points) reflect different levels of performance on the question.

Slide 25

- Without getting into the technical details of measurement error here, let's just say that the less measurement error there is, the greater the test reliability is. And there are many factors that contribute to measurement error … scoring error is just one of them.
- Yes, we can award zero points or one point to the answer options a student chooses on a multiple-choice test, and do that perfectly every time. But don't forget, reliability and content validity are largely a function of the quality of the sampling of the content and skills of interest. The more items and the better content representation, the greater the reliability and content validity -- regardless of the item type.

Slide 26

- It's unfortunate that multiple-choice tests are considered objective measures and that human scoring of student work is considered subjective.
- For large-scale testing programs, many practices are designed to make the human scoring process as objective as possible. Scorers have to have the proper credentials, they are trained to use scoring rubrics on sample responses, they have to qualify to score each item they score, their accuracy is continuously monitored, and they don't know the names of the students or schools.
- There is some scoring error, but think about it…if a response got 3 points when it should have gotten the maximum 4 points, it still got more than half credit. Some error is acceptable. After all, if a student can get a point for guessing correctly on a multiple-choice question, we accept that error. Reliability and validity apply to whole tests.

Slide 27

- In scoring his or her own students' work, a teacher should apply the same rules consistently for all students. But the teacher does not have to score student work the same way other teachers would.

- The exception would be when different teachers teach the same course to different students in the same school. Then common tests and scoring criteria, and grading practices for that matter, would be advisable in the interest of fairness.

Slide 28

<blank>

Slide 29

- It should be clear that item types vary in terms of the information they yield on student learning. That information is not just important for purposes of reliability and validity. It is also important for instructional purposes and for understanding where students stand relative to their understanding of curriculum content and skills.
- A recommended activity, which you will find in the Lesson 4 Supplement, illustrates this point further.

Slide 30

- Teachers . . . use a balance of item types throughout the school year and in your bigger tests.
- Don't fall into the trap of over reliance on multiple-choice tests because of their convenience. They work well for low level skills, but not for deeper learning.
- Make use of resources such as textbooks on classroom assessment to learn how to develop effective items, tasks, and scoring rubrics, and to learn various techniques for efficient and accurate scoring of student work on constructed-response and performance tasks.

Slide 31

- Using a set of survey items dealing with facts or beliefs about different kinds of test items, this activity calls for a discussion with a colleague about the survey statements. The discussion should involve points made in this lesson.

Slide 32

- This Scantron blog post shows the relationship between item type and cognitive complexity, referring to three different models of complexity. Through a simple graphic, it also makes a good case for the use of multiple item types in testing.
- The Lesson 4 Supplement also identifies the same two books listed for Lesson 3 as useful references on student testing for teachers to use throughout their careers.