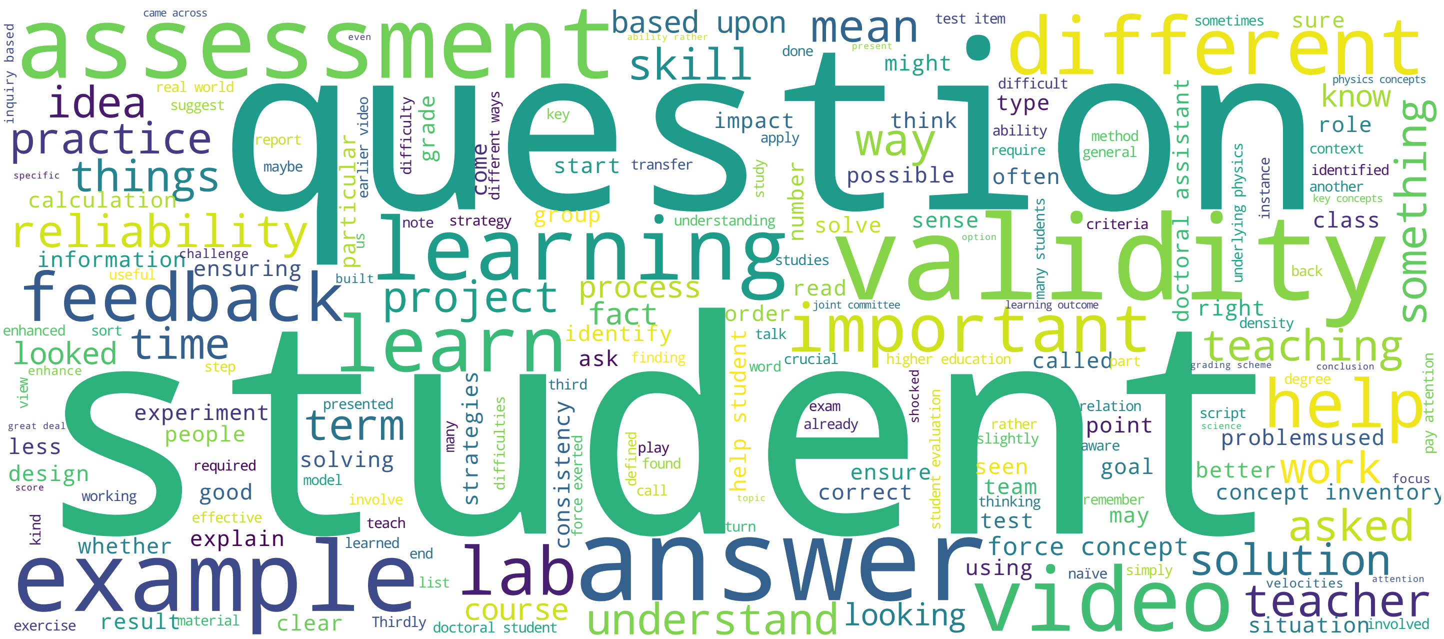
Valid and Reliable Assessment of Learning

Foundations for Teaching in Higher Education

Roland Tormey
Ingrid Le Duc
Cécile Hardebolle

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE



Search MOOC

Video

EPFL

Assessment is in many respects the Cinderella of teaching and learning in higher education. Many books on teaching and learning in higher education don't even include a chapter on assessment and yet, assessment is crucial. It's crucial for two reasons; first of all, universities play a certification role. We certify that our students do in fact have the skills that we've tried to teach them. Many of our students will go out having the right to call themselves engineers, they will be certified by accrediting bodies who have given us the right to give them that title and therefore, we need to be able to ensure that they do in fact have the skills that we claim they have. The second role for assessment in higher education is maybe even more important. Assessment has been called by one prominent educational researcher 'the tail that wags the curriculum dog'. Oftentimes, when students want to understand what it is that we really want them to learn, they look to the assessment.

Notes

Summary

0m 04s

Goals of this video

You should be able to:

- Define what is meant by **validity** in relation to assessment
- Define what is meant by **reliability** in relation to student assessment

So in this video, we're concerned with two key concepts, which apply to assessment. The first of these is validity; we'll want you to understand what validity means and what it is that a doctoral system can do to enhance the validity in assessment. Secondly the concept is reliability and again, we want you to understand what that means in relation to educational assessment and how you can enhance the reliability of assessment.

Notes

Summary

1m 07s

What is validity?

Eric Mazur (Harvard)

- Force Concept Inventory (FCI)
  - Physics without calculation

So let's take the first topic, what is validity? To do this, I'm going to start with a practical example. This example is provided by Eric Mazur who's a professor of physics in Harvard. Mazur explains that in the late '80s early '90s, he came across an instrument, a test called the force concept inventory In fact, we've touched on this test earlier in one of our earlier videos. The force concept inventory is a test, which is designed to test whether or not students understand how physics concepts apply in the real world settings. So they include test items which involved no calculation whatsoever. An example of a test item would be there is a large truck and a small car, they push against each other. Is the force exerted by the car on the truck greater than, less than, or the same as the force exerted by the truck on the car. This force concept inventory includes elements, which on the surface of it may seem relatively simple and certainly, that was what Mazur thought when he first came across it. But he was struck by the finding that many students who could pass first-year physics actually struggled with these test items.

Notes

Summary

1m 36s

What is validity?

Eric Mazur (Harvard)

- Force Concept Inventory (FCI)
  - Physics without calculation
- 2 similar problems on a mid-term exam
  - Conceptual version: avg. 4.9
  - Conventional version: avg. 6.9

His initial response was to say well "Not my students!" and so therefore, he took the test and he administered it to his students. He discovered that in fact his students did not perform significantly better on the force concept inventory than they had performed on their midterm exam. This shocked him because the force concept inventory questions appeared to be a good deal simpler than his questions in his exam, which required significant amounts of calculation. So he decided to explore this slightly further. On a future midterm, he gave the students five questions two of which were actually the same question albeit in a different form. In question one, it was presented as a conceptual question with no calculation. In question five it was presented as a calculation question. The format of the questions look different so it wouldn't be immediately obvious to someone looking at the diagrams that these were the same question. Nonetheless, the underlying physics concepts tested in these two questions were the same. And he was shocked by the results. His students performed better on the conventional version of the question where they had to do some calculations, than they did with a conceptual version of the question, which required no calculations.

Notes

Summary

3m 02s

What is validity?

Eric Mazur (Harvard)

- Force Concept Inventory (FCI)
  - Physics without calculation
- 2 similar problems on a mid-term exam
  - Conceptual version: avg. 4.9
  - Conventional version: avg. 6.9
- "it is possible for students to do well in conventional problems by memorizing algorithms without understanding the underlying physics" (Mazur, 1997, p. 6)

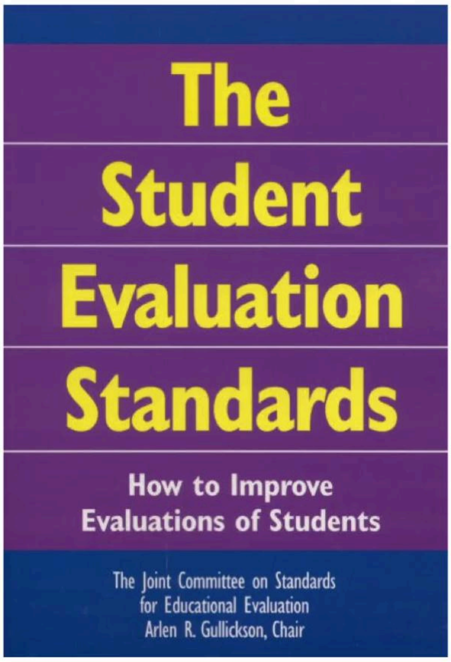
The apparently easier question turned out to be more difficult than the standard university-level questions. Now, when we looked at this in an earlier video, we looked at the idea that students bring with them naïve or intuitive scientific views, which they struggle to give up during their physics learning and during their scientific learning in university. The questions which were in the force concept inventory were specifically designed to unearth these naïve or non-scientific worldviews, which explains what it is that students perform less well. But in a sense it means more than that because what Mazur concluded from this is that it's possible for students to do well in conventional problems by memorizing algorithms without actually understanding the underlying physics. In other words, it was possible for students to do well in a conventional physics exam without actually understanding the underlying physics. So this brings us to the question of validity, this helps to explain what we mean by validity.

Notes

Summary

4m 27s

What is validity?



"Validity refers to the degree to which inferences drawn from the results of the assessment….are trustworthy"

"It is incorrect to say that a specific assessment method is valid. Rather it is the inferences drawn from the information gained… that must be valid"

"Validity is the **single most important issue** in student evaluation" (emphasis added)

(Gullickson, 2003, p. 127-128).

So Mazur has given us an example of validity in practice or 'in-validity' in practice we might say. How is validity actually defined? So to get a definition of validity, we can turn to this resource; 'the student evaluation standards', which is the result of a joint committee, which was put together by a number of different educational and psychological testing organizations in order to identify what are the minimum standards that should be applied whenever we assess students. So the student evaluation standards define validity as follows; validity refers to the degree to which inferences drawn from the results of an assessment are trustworthy. So it's important to note in that, that it's actually not the assessment itself which is valid, it's the inferences which are drawn from it. For the Joint Committee on standards for educational evaluation, validity was the single most important issue in student evaluation. If there was one thing to get right, it was validity.

Notes

Summary

5m 37s

Validity in practice

- "Contextualize theoretical models in every day life"
  - Do the questions ask student to do this?
  - Do you encourage students to see this as important?
  - Does the grading scheme reward them for doing this?

So what can a doctoral assistant do in practice to helping improve the validity of an assessment. Well one thing is to be clear as to what the actual goals of the course are. What is it that the teacher wants the students to learn? This might seem obvious but sometimes, it's not all that obvious, let's take some examples. Here's a learning outcome which is listed on lots of different physics courses "contextualize theoretical models in everyday life". So an obvious question for validity is whether or not students are actually asked to do this in the test questions which they do. But a second important point in validity is whether or not the students are actually aware that this is one of the things that they're going to be asked to do. Students will often pay attention to the derivations and the formulas in a physics class and when the teacher starts to talk about how that applies in real-world life, students will often think well "okay this is less important, this is just an example". If it's a goal of the course that the student is able to do this transfer, well then the student needs to have their attention directed to that in order that they pay attention to the things that they actually need to learn.

Notes

Summary

6m 43s

Validity in practice

- "Contextualize theoretical models in every day life"
  - Do the questions ask student to do this?
  - Do you encourage students to see this as important?
  - Does the grading scheme reward them for doing this?
- "Design an experiment to test a hypothesis relating to …."
  - Does the assessment task ask them to do this?
  - Does the grading scheme reward them for doing this?

Obviously, if this is the goal of the course and if it's built into the assessment, then it also needs to be built into the grading scheme in some way. A second example we might think about is designing an experiment to test a hypothesis relating to the material in question. This is a learning outcome, which is often used in relation to lab courses and again, the same kind of principles employee. Does the assessment of the lab actually asked them to do this? Are they ever asked to design an experiment? If they are not asked to design an experiment, well then that raises questions about the validity of the assessment. When they write the report and when that's being graded, are they getting specific marks for doing this? Are they being rewarded for actually showing that they've learned the things that are identified as the learning goals of the course?

Notes

Summary

7m 52s

What is reliability?

"Reliability refers to the degree of consistency of the scores or information obtained from an information-gathering process" (Gullickson, 2003, p 161)

- Would the student obtain the same score if two different teachers graded the exam?
- ...or if the same teacher graded the exam at two different times?

So in this video we're looking at two ideas the first of which is validity, the second of which is reliability. So what does reliability mean? Well reliability is defined as the degree of consistency of the scores or information obtained from an information gathering process. That’s a kind of an abstract definition. What does it mean in practice? Well let's take some examples, let's imagine first of all a student in a big course. There are many students taking the course and therefore, the scripts are divided between for example different doctoral assistants. How sure are we that the student will get the same grade if their script is given to doctoral student A or doctoral student B? that is a question of consistency, that is a question of reliability. Let’s think of another example; how sure are we that the student will get the same grade if the teacher grades their script on a Wednesday morning and again on a Friday evening? Again, this is a question of consistency, this is a question of reliability.

Notes

Summary

8m 48s

Improving test reliability (1)

Ball A with mass m and velocity v_0 is moving along the same straight line as ball B with mass M and velocity u_0. The two balls collide. The coefficient of restitution between the balls is e.

Express the velocities of the two balls after the impact in terms of their velocities before the impact.

So what can be done in practice to enhance test reliability? So let's look at an example here of a question which can be given to a student. I'm going to suggest you pause the video now to take a look at this question and just read through it for a moment before continuing.

Notes

Summary

9m 58s

Improving test reliability (1)

Ball A with mass m and velocity $\boldsymbol{v_0}$ is moving along the same straight line as ball B with mass M and velocity $\boldsymbol{u_0}$. The two balls collide. The coefficient of restitution between the balls is e.

Express the velocities of the two balls after the impact in terms of their velocities before the impact.

What's the answer?

$$v_0\boldsymbol{i} - u_o\boldsymbol{i} = -e(v_1\boldsymbol{i} - u_1\boldsymbol{i})$$

$$\boldsymbol{v_1} = (\frac{(m - eM)v_o + M(1+e)u_o}{m+M})\boldsymbol{i}$$

$$\boldsymbol{u_1} = (\frac{m(1+e)v_o + (M - em)u_o}{m+M})\boldsymbol{i}$$

So what's the answer to this question? Well in actual fact, there are multiple answers to this question and this is going to cause a difficulty from the point of view of reliability. here's one answer to the question, it's a very straightforward answer, it doesn't require a lot of thought, simply the student has to identify what coefficient of restitution means and basically they have to write down in a formula terms what that means in practice. So this is just about recognition and writing down. It's correct, but it's not a very deep answer. This is the second answer to the question, which actually involves a great deal more working out. Again, this answer is correct but it's an answer which is required a great deal more effort from the point of view of the student. So which answer here is the one that the teacher is looking for? Is it the first answer, is it the second answer? If the teacher is looking for the first answer then a student who works out the second answer has actually wasted time on this question that they could better spend elsewhere; they've lost time. If the answer the teacher is looking for is the second answer, well then actually the student has - through no error of their own - failed to respond to the question the teachers asked.

Notes

Summary

10m 15s

Improving test reliability (1)

Ball A with mass m and velocity $\boldsymbol{v_0}$ is moving along the same straight line as ball B with mass M and velocity $\boldsymbol{u_0}$. The two balls collide. The coefficient of restitution between the balls is e.

Express the velocities of the two balls after the impact in terms of their velocities before the impact.

What's the answer?

$$v_0\boldsymbol{i} - u_o\boldsymbol{i} = -e(v_1\boldsymbol{i} - u_1\boldsymbol{i})$$

$$\boldsymbol{v_1} = (\frac{(m - eM)v_o + M(1+e)u_o}{m+M})\boldsymbol{i}$$

$$\boldsymbol{u_1} = (\frac{m(1+e)v_o + (M - em)u_o}{m+M})\boldsymbol{i}$$

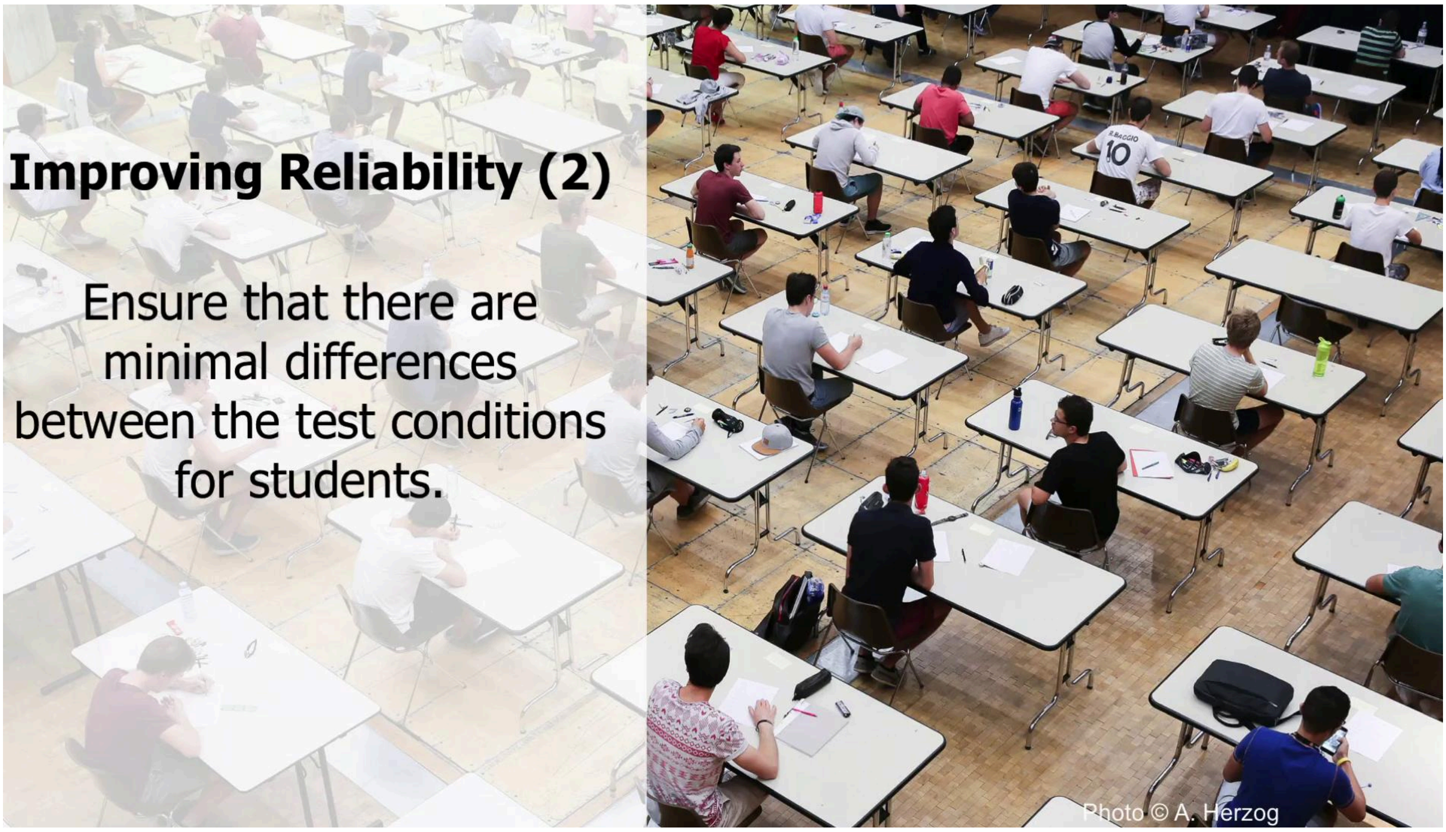
"For both ball A and ball B, express its velocity after the impact in terms of m, M, e, v_0 and u_0".

This is a situation where students might get grades not based upon their actual ability but based upon whether or not they read the question in the sense that the teacher intended it, rather than simply in the sense in which it was written. A solution here would be to rephrase the question in a slightly more expressive way. When the student is asked to express the velocities of the two balls after the impact in terms of the different components, m and M and their velocities before the impact. This will make clear to the student what form the answer is to come in, and it will ensure that all students will be judged based upon their ability rather than based upon luck.

Notes

Summary

11m 33s



A second way in which we can improve reliability is by ensuring that students are assessed in more or less the same test conditions. Again, this is something where a doctoral student will very often be very important because very often it will be the doctoral assistant's job to supervisor or invigilate the exams to ensure that there are similarity in conditions across students.

Notes

Summary

12m 17s

Conclusion

- Validity
  - Test what you want them to learn
  - Communicate to them what's important
  - Reward these things in grading
- Reliability
  - Clarity in questions and instructions
  - Ensure similarity of treatment
  - Have a system to ensure consistency

So in this video, we've looked at two of the key concepts which are important in assessment. The first of these is validity; as a doctoral assistant you can play a key role in validity in a number of different ways. First of all, validity will be enhanced when we test what it is that we want the student to learn. Secondly, it will be enhanced when the student is aware of what it is that we want them to learn and what it is they're going to be tested on. Thirdly, it will be enhanced when the grading scheme actually rewards the student for doing the things that we've said we want them to learn. The second concept we looked at was reliability and reliability can be improved in an of different ways. In this video we've looked at two. First of all, clarity in terms of questions and instructions can be really important in terms of ensuring reliability, in terms of ensuring that students grade is based upon their ability rather than based upon look. Secondly, ensuring an equality of treatment of students during assessment is an important role and it's very often a role the doctoral assistants will play. At third way of enhancing reliability is to have a system in place to ensure some sort of consistency between different graders or in the same grader over time. And that is something that we look at in the next video.

Notes

Summary

12m 42s