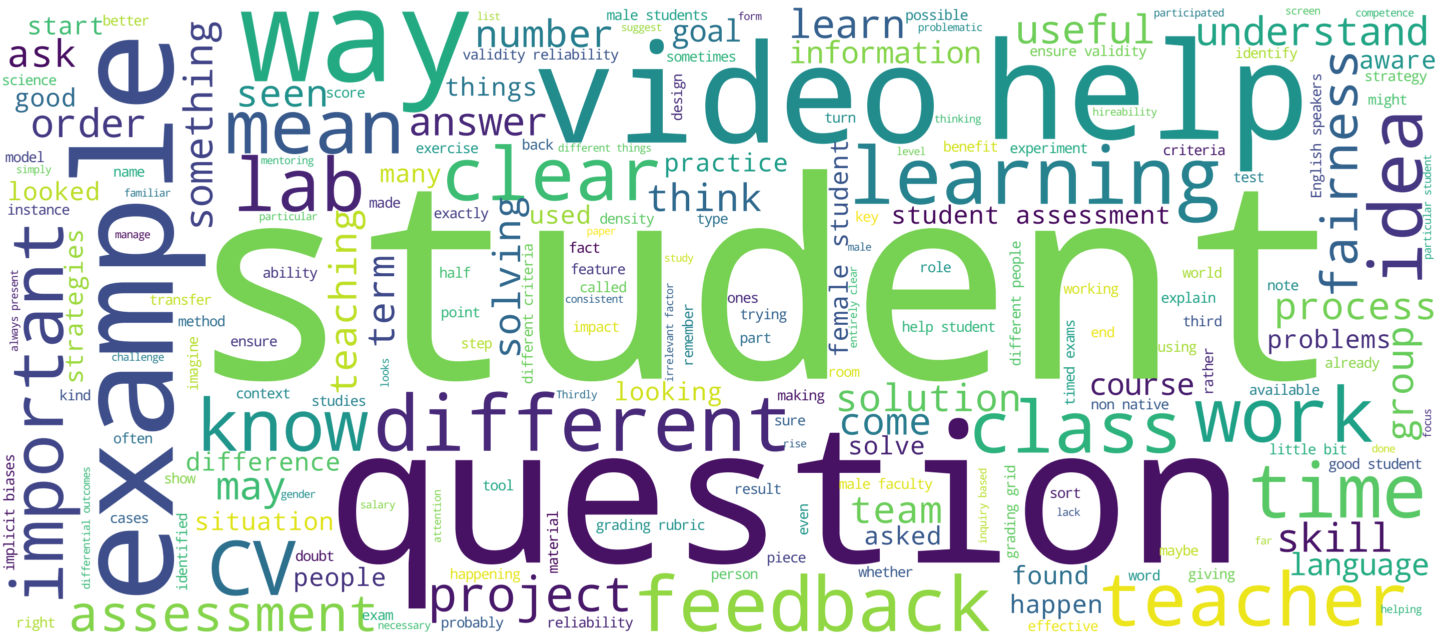
Valid and Reliable Assessment of Learning

Foundations for Teaching in Higher Education

Roland Tormey
Ingrid Le Duc
Cécile Hardebolle
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE



Search MOOC

Video

EPFL

Goals of this video

You should be able to:

- Define what is meant by **fairness** in relation to assessment
- Identify the role a grading rubric can play in maximizing reliability, validity and fairness in assessment

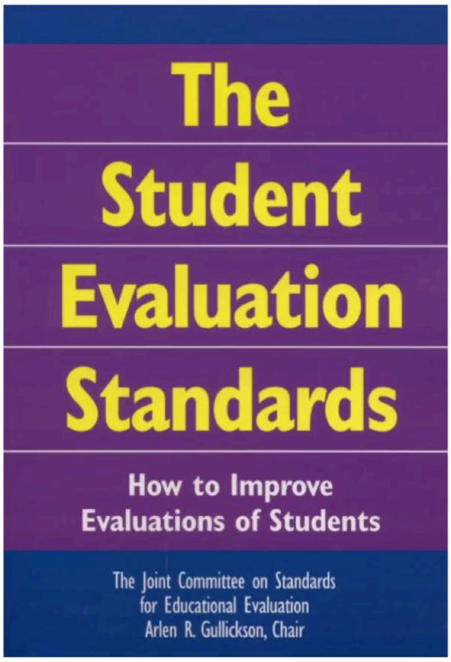
So this is the second of two videos which looks at the question of assessment. In the first video, we looked at the idea of validity and reliability and we looked at what they mean in practice for doctoral assistants. In this video, we're going to look at the idea of fairness and we're also going to look at a tool, which is useful in helping to ensure validity, reliability and fairness in student assessment. So the goals for this video are first of all, that the idea of 'fairness' will make sense to you; that you'll be able to define what fairness means in the context of student assessment. Secondly, we want to look at a particular tool, which is called a grading rubric and the look at the way in which grading rubrics can help ensure validity, reliability and fairness in student assessment.

Notes

Summary

0m 04s

What is fairness or bias in assessment?



Fairness and Bias

"Bias occurs when irrelevant or arbitrary factors systematically influence interpretations and judgements… in a way that affects… subgroups of students"

(Gullickson, 2003, p.167)

In relation to student assessment, we can think of fairness is meaning and absence of bias. So what is bias? Well in the student assessment, bias occurs when irrelevant or arbitrary factors systematically influence our interpretations in a way that effects particular students or particular subgroups of students. Now, that's our formal definition. What does that mean in in practice?

Notes

Summary

0m 51s

Bias in practice

Individual bias

- "He was always in class and participated well – I know what he meant in this case even if he didn't say it very clearly"

Well let's imagine a particular situation, let's say that when you're grading the work of students, you come to the work of a particular student, a student that you happen to know because this student was always present in class, they participated well in discussions. When you go through that work and maybe it's not entirely clear on some occasion what a student means or it's not entirely clear if a student has understood something well, do you find yourself giving them the benefit of the doubt? Do you find yourself saying well this student was always present in class, they participated well in these questions, I've seen from them already that they understand this material so therefore I will give them the benefit of the doubt in this case. Do you also extend that same benefit of the doubt to other students, students who never turned up in class. If you don't, then there's a bias creeping in, an irrelevant factor that is whether or not the student turned up in class, is influencing your judgment as to what is the student skill. And so that's an example of individual bias. Biases don't just happen on an individual basis and they don't just happen by intention.

Notes

Summary

1m 15s

Bias in practice

Individual bias

- "He was always in class and participated well – I know what he meant in this case even if he didn't say it very clearly"

Language bias

- "A mass *m* is attached to a thin (massless) cable and is whirled around in the vertical plane at a constant velocity..."
- Especially in timed examinations

Sometimes biases can be seen in the differential outcomes that come for students, not because of the intention of the teacher. Let's look at an example of this case, this could be an example of a language bias. So imagine a question which starts off 'a mass M is attached to a thin massless cable and is whirled around in a vertical plane at a constant velocity...'. this question here has a word in it 'whirled', which may be very familiar to English speakers but may not be all that familiar to non-native English speakers. What does that mean? It means that non-native English speakers will take longer to understand what's actually happening in this question. It may be that they'll understand very well what's happening in a question but it will take them time. And so therefore, forms of words, which are not widely known to non-native speakers may pose specific difficulties particularly in the cases of timed exams. In a more general sense, timed exams can be problematic where you have some people in the room for whom the language is their native language and some people in the room for whom the language is a second language.

Notes

Summary

2m 23s

Bias in practice

Individual bias

- "He was always in class and participated well – I know what he meant in this case even if he didn't say it very clearly"

Language bias

- "A mass *m* is attached to a thin (massless) cable and is whirled around in the vertical plane at a constant velocity…"
- Especially in timed examinations

even if they are very fluent in that second language, it will probably take them a little bit longer to understand what's going on and formulate what they're trying to say in that second language. Therefore, timed exams can be problematic in this case, it can introduce a bias.

Notes

Summary

3m 41s

Gender bias in undergraduate science education

- Corinne Moss-Racusin et al. (2012)
- Yale Universty

- The study:
  - Good (not excellent) application materials of student applying for a lab manager position
  - Same application, randomly described as John or Jennifer
  - Scored by 127 biology, chemistry and physics professors


Science faculty's subtle gender biases favor male students

Corinne A. Moss-Racusin[a,b], John F. Dovidio[b], Victoria L. Brescoll[c], Mark J. Graham[a,d], and Jo Handelsman[a,1]

[a]Department of Molecular, Cellular and Developmental Biology, [b]Department of Psychology, [c]School of Management, and [d]Department of Psychiatry, Yale University, New Haven, CT 06520

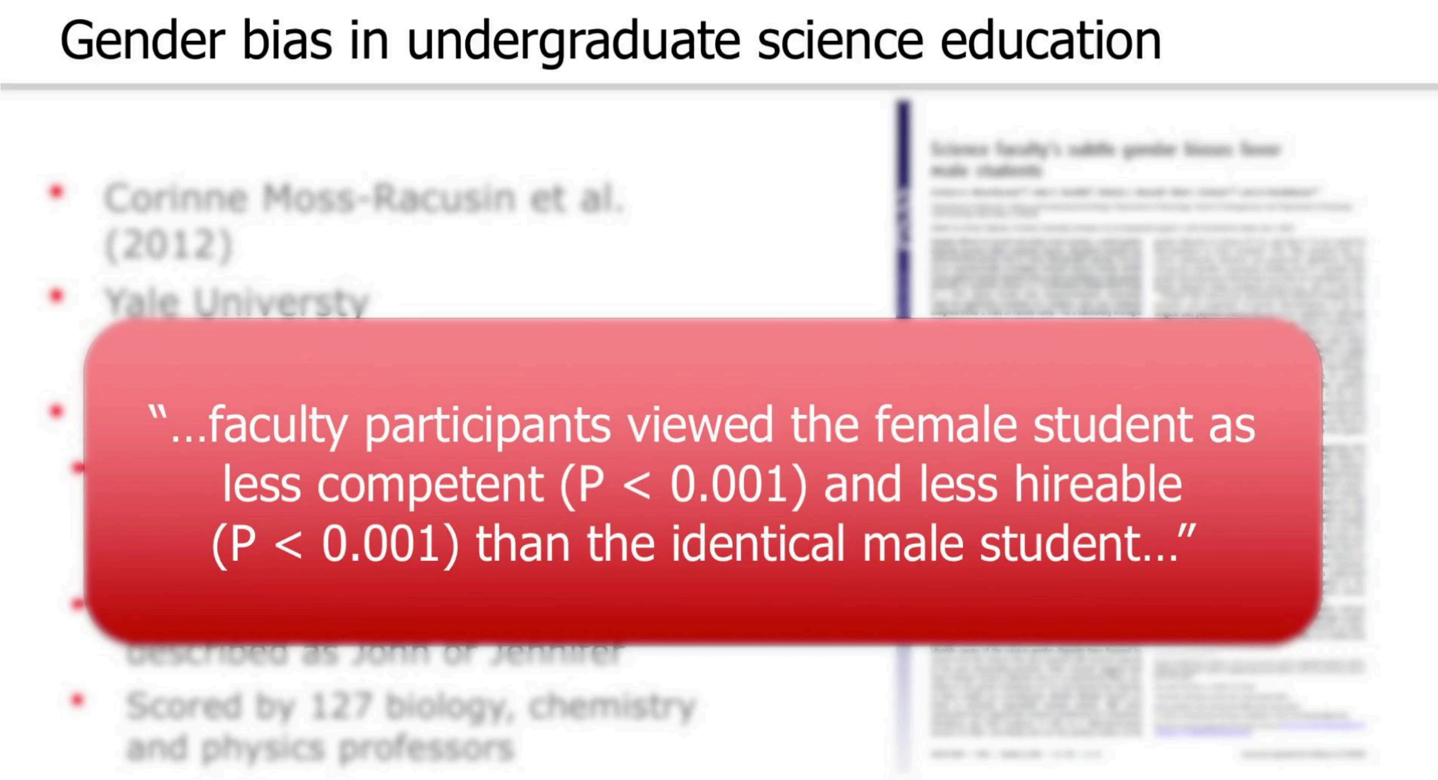
Edited* by Shirley Tilghman, Princeton University, Princeton, NJ, and approved August 21, 2012 (received for review July 2, 2012)

Another example of perhaps unintentional bias is gender bias, which sometimes arises in science education. So to look at this I'm going to use this paper here which is published in 2012 by Corrine Moss-Racusin from Yale University in which she and her colleagues looked at gender bias in science faculty. Their method of working here was that they took a CV of a good student; not an excellent student, a good student. And they distributed the CV to 127 different science faculty; biology, chemistry, physics professors, and they asked them to rate the CV on a number of different criteria. What made the study notable was that in half of the cases the CV had the name 'John' on it and in the other half, the CV had the name 'Jennifer' on it. there was no other difference between the CV's, the only difference was the name which identified implicitly the gender of the person who was submitting the CV.

Notes

Summary

3m 58s



What they found is that when the name on the CV was Jennifer, faculty participants saw the person who had submitted the CV as being less competent and less hireable than when the name on the CV was male.

Notes

Summary

5m 07s

A battle of the sexes?

Scoring of Male and Female Faculty of Applications

	"Male" Student		"Female" student	
	Male Faculty	Female faculty	Male Faculty	Female faculty
Competence	4.01	4.1	3.33	3.32
Hireability	3.74	3.92	2.96	2.84
Mentoring	4.74	4.73	4.00	3.91
Salary	30, 520	29,333	27,111	25,000

One might hypothesize that there's some sort of bias by male faculty members against female students here but in actual fact, it's useful therefore to look in a little bit more detail at the actual data that they found.

Notes

Summary

5m 24s

A battle of the sexes?

Scoring of Male and Female Faculty of Applications

	"Male" Student		"Female" student	
	Male Faculty	Female faculty	Male Faculty	Female faculty
Competence	4.01	4.1	3.33	3.32
Hireability	3.74	3.92	2.96	2.84
Mentoring	4.74	4.73	4.00	3.91
Salary	30, 520	29,333	27,111	25,000

So in this case, we have the scores by male faculty for both male students on the left of the screen and female students on the right of the screen. What we see is that male faculty did in fact score the male students higher for competence for hireability for mentoring and for salary than they did the female students.

Notes

Summary

5m 37s

A battle of the sexes?

Scoring of Male and Female Faculty of Applications

	"Male" Student		"Female" student	
	Male Faculty	Female faculty	Male Faculty	Female faculty
Competence	4.01	4.1	3.33	3.32
Hireability	3.74	3.92	2.96	2.84
Mentoring	4.74	4.73	4.00	3.91
Salary	30, 520	29,333	27,111	25,000

However, when we look at female faculty, we see that actually they do the same thing. Male students were scored higher for competence, for hireability, for mentoring and for salary than were the female students by the female faculty.

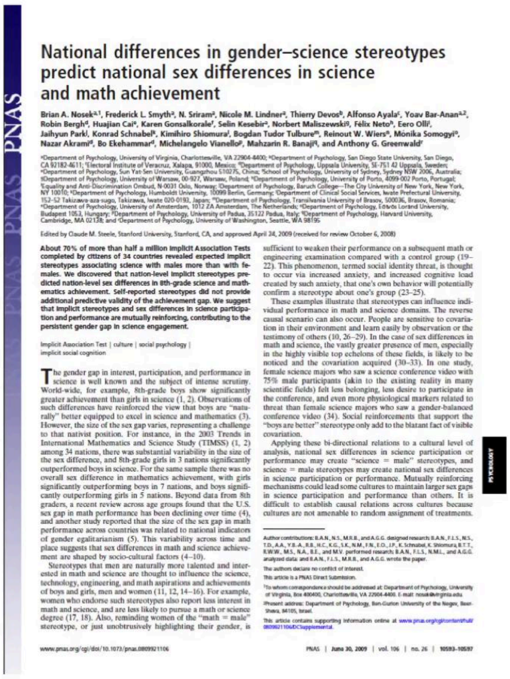
Notes

Summary

5m 59s

Unconscious bias

- People are often unaware of their own implicit biases
- These may not align with people's worldview
- Have been found to be associated with gender gap in science and math attainment (Nosek, 2009)


PNAS

National differences in gender–science stereotypes predict national sex differences in science and math achievement

Brian A. Nosek[a,1], Frederick L. Smyth[a], N. Sriram[a], Nicole M. Lindner[a], Thierry Devos[b], Alfonso Ayala[c], Yoav Bar-Anan[a,2], Robin Bergh[d], Huajian Cai[e], Karen Gonsalkorale[f], Selin Kesebir[a], Norbert Maliszewski[g], Félix Neto[h], Eero Olli[i], Jaihyun Park[j], Konrad Schnabel[k], Kimihiro Shiomura[l], Bogdan Tudor Tulbure[m], Reinout W. Wiers[n], Mónika Somogyi[o], Nazar Akrami[d], Bo Ekehammar[d], Michelangelo Vianello[p], Mahzarin R. Banaji[q], and Anthony G. Greenwald[r]

[a]Department of Psychology, University of Virginia, Charlottesville, VA 22904-4400; [b]Department of Psychology, San Diego State University, San Diego, CA 92182-4611; [c]Electoral Institute of Veracruz, Xalapa, 91000, Mexico; [d]Department of Psychology, Uppsala University, SE-751 42 Uppsala, Sweden; [e]Department of Psychology, Sun Yat-Sen University, Guangzhou 510275, China; [f]School of Psychology, University of Sydney, Sydney NSW 2006, Australia; [g]Department of Psychology, University of Warsaw, 00-927, Warsaw, Poland; [h]Department of Psychology, University of Porto, 4099-002 Porto, Portugal; [i]Equality and Anti-Discrimination Ombud, N-0031 Oslo, Norway; [j]Department of Psychology, Baruch College—The City University of New York, New York, NY 10010; [k]Department of Psychology, Humboldt University, 10099 Berlin, Germany; [l]Department of Clinical Social Services, Iwate Prefectural University, 152-52 Takizawa-aza-sugo, Takizawa, Iwate 020-0193, Japan; [m]Department of Psychology, Transilvania University of Brasov, 500036, Brasov, Romania; [n]Department of Psychology, University of Amsterdam, 1012 ZA Amsterdam, The Netherlands; [o]Department of Psychology, Eötvös Loránd University, Budapest 1053, Hungary; [p]Department of Psychology, University of Padua, 35122 Padua, Italy; [q]Department of Psychology, Harvard University, Cambridge, MA 02138; and [r]Department of Psychology, University of Washington, Seattle, WA 98195

Edited by Claude M. Steele, Stanford University, Stanford, CA, and approved April 24, 2009 (received for review October 6, 2008)

About 70% of more than half a million Implicit Association Tests completed by citizens of 34 countries revealed expected implicit stereotypes associating science with males more than with females. We discovered that nation-level implicit stereotypes predicted nation-level sex differences in 8th-grade science and mathematics achievement. Self-reported stereotypes did not provide additional predictive validity of the achievement gap. We suggest that implicit stereotypes and sex differences in science participation and performance are mutually reinforcing, contributing to the persistent gender gap in science engagement.

Implicit Association Test | culture | social psychology | implicit social cognition

The gender gap in interest, participation, and performance in science is well known and the subject of intense scrutiny. World-wide, for example, 8th-grade boys show significantly greater achievement than girls in science (1, 2). Observations of such differences have reinforced the view that boys are "naturally" better equipped to excel in science and mathematics (3). However, the size of the sex gap varies, representing a challenge to that nativist position. For instance, in the 2003 Trends in International Mathematics and Science Study (TIMSS) (1, 2) among 34 nations, there was substantial variability in the size of the sex difference, and 8th-grade girls in 3 nations significantly outperformed boys in science. For the same sample there was no overall sex difference in mathematics achievement, with girls …

It's important to note the people may not themselves be actually aware of these biases. It's not something that they consciously hold in their mind; a belief that women are less capable in science than men. Rather it's something which is implicit, which they have absorbed from their culture and which they themselves may not even be aware of. In fact, studies have looked at this all across the world and have identified that there are quite big cultural differences between the implicit biases which are found in different cultures. They've also identified that these differences in implicit biases are actually correlated with differences in performance. In other words, countries which show a greater degree of implicit bias are the same countries in which the gap between male and female performance on science and math tends to be greatest.

Notes

Summary

6m 13s

Procedures to ensure validity, reliability, fairness

Reliability & Fairness

- Know what your criteria of assessment are.
- Know what excellent, good, and good enough performance looks like?
- Ensure everyone marking scripts has the same criteria and guidelines.

Validity

- Ensure students have the same guidelines

So how do you maximize fairness in assessment of students? Well sometimes, it's possible to be aware of features which give rise to a lack of fairness: language is an example in this case. So sometimes, it's necessary to set up the assessment to make sure that things like language don't actually get in the way and produce unfairness in the results. However oftentimes, we can be unaware of the features which are giving rise to a lack of fairness. For example in the case of implicit biases. In this kind of situation, then having clearly documented procedures which we follow will be important in ensuring that students get the outcome that they deserve. So it can be very important to ensure that in order to maximize fairness and indeed reliability, that we are really clear about what the criteria of assessment are, what exactly we are looking for students to do in an assessment. Secondly, that we are also clear about what we mean by excellent, good and good enough. What are our key benchmarks for how good a student should be at a particular piece of work. Obviously, if there are different people involved in assessing the student, it's important that they all have the same criteria.

Notes

Summary

7m 07s

Procedures to ensure validity, reliability, fairness

Reliability & Fairness

- Know what your criteria of assessment are.
- Know what excellent, good, and good enough performance looks like?
- Ensure everyone marking scripts has the same criteria and guidelines.

Validity

- Ensure students have the same guidelines

From the students point of view, these criteria can also play a role in ensuring validity. If the student is clear about what it is that we want them to do, then good students will go and do it and less good students will not manage to do it. If the student is unclear about what it is we want them to do, then lucky students will do it by accident and unlucky students will fail to do it. in this case, giving students a clear account of what it is we're looking for from them will help to ensure that the students who are good will be the ones who will score well rather than simply the ones who are lucky.

Notes

Summary

8m 22s

Improving test reliability using a rubric

Heading	Description	Marking
Conceptual Understanding	Correctly identifies the concepts / formulae / approaches that are appropriate for solving the problem. Marks may be reduced if only part of the conceptual basis of a question or only some relevant mathematical elements are correctly identified.	Up to 0.4 marks
Calculation and Derivation	Calculates or derives answers with accuracy, arriving at the correct answer (within ±1% for problems with numerical values; within ±5% for graph-based problems). Marks may be reduced for calculation errors or errors in derivation. Correct calculations which are based on incorrect concepts/ formulae will score zero.	Up to 0.5 marks
Communication	The written work clearly communicates how the student arrived at the solution. Well communicated incorrect work will score zero.	Up to 0.1 marks

So one tool that is sometimes used for ensuring validity, reliability and fairness in assessment is a grading rubric or a grading grid. So in this case here, we have a grid, which makes a clear to the student and indeed to ourselves what are our criteria.

Notes

Summary

8m 59s

Improving test reliability using a rubric

Heading	Description	Marking
Conceptual Understanding	Correctly identifies the concepts / formulae / approaches that are appropriate for solving the problem. Marks may be reduced if only part of the conceptual basis of a question or only some relevant mathematical elements are correctly identified.	Up to 0.4 marks
Calculation and Derivation	Calculates or derives answers with accuracy, arriving at the correct answer (within ±1% for problems with numerical values; within ±5% for graph-based problems). Marks may be reduced for calculation errors or errors in derivation. Correct calculations which are based on incorrect concepts/ formulae will score zero.	Up to 0.5 marks
Communication	The written work clearly communicates how the student arrived at the solution. Well communicated incorrect work will score zero.	Up to 0.1 marks

So in this case, it distinguishes between conceptual understanding calculation & derivation and communication. These are the three different grading criteria. It also makes clear how much marks is being awarded for each of those three different criteria. Within the description, it then makes clear what's the distinction between a really good performance and a good enough performance.

Notes

Summary

9m 18s

Improving test reliability using a rubric

Heading	Description	Marking
Conceptual Understanding	Correctly identifies the concepts / formulae / approaches that are appropriate for solving the problem. Marks may be reduced if only part of the conceptual basis of a question or only some relevant mathematical elements are correctly identified.	Up to 0.4 marks
Calculation and Derivation	Calculates or derives answers with accuracy, arriving at the correct answer (within ±1% for problems with numerical values; within ±5% for graph-based problems). Marks may be reduced for calculation errors or errors in derivation. Correct calculations which are based on incorrect concepts/ formulae will score zero.	Up to 0.5 marks
Communication	The written work clearly communicates how the student arrived at the solution. Well communicated incorrect work will score zero.	Up to 0.1 marks

It clarifies what will you gain you extra marks and what will lose you extra marks. Having a grading grid like this, making it available to all of the assessors and indeed making it available to the students in advance of the exam will help to ensure validity, reliability and fairness. So if you're in a course and you're not all that clear as to what the criteria are, you're not all that clear as to what really 'excellent' looks like and what 'good enough' looks like, it might be useful to actually sit down with the teacher and for yourself help to map out what is the grading grid for the course.

Notes

Summary

9m 40s

Conclusion

- Validity
  - Test what you want them to learn
- Reliability
  - Ensure consistency
- Fairness
  - Ensure irrelevant factors do not interfere
- Document your grading criteria
  - Share it with other examiners
  - Share it with students

So when we look at assessment there are a number of different things to take into account. first of all, validity; validity is perhaps the most important criteria. We want to test what it is that we want the student to learn and we want the student to know what it is that we're going to test them on. Secondly, reliability; you want to ensure that we're consistent whether that's consistent in terms of me marking the same piece of work at different times or whether that's consistent in terms of different people marking students' work. Ensuring that consistency across markers or across time will be crucially important. Thirdly, we want to avoid bias; we want to ensure fairness in as far as possible. we want to ensure that irrelevant factors like for example language or gender don't get in the way and give rise to differential outcomes for different people. in order to do these different things, it can be really useful to document your grading criteria, share it with other examiner's to ensure reliability and fairness and also, share it with students in order to help to maximize validity.

Notes

Summary

10m 17s