



UNIVERSITY *of* CAMBRIDGE
International Examinations

Principles of Assessment

Nicki Little





4 Principles of Assessment

- Validity
- Reliability
- Fairness
- Manageability / practicality



Validity – some definitions

- A test is valid if it gives an accurate measurement of what it is intended to measure
- ‘The fitness for purpose of an assessment tool or scheme’ (QCA)
- ‘The students’ minds are doing the things we want them to show us they can (or can’t) do.’



Validity – another definition

- A measure is valid if it does what it is intended to do – which is typically to act as an indicator of an abstract concept (for example, height, weight, time etc.) which it is intended to measure.

(Dictionary of Language Testing)



Validity - a matter of interpretation

- ‘The extent to which the inferences or decisions we make on the basis of test scores are meaningful, appropriate and useful’
(Bachman / APA)
- What the results of our assessment mean
(Wiliam)



Six kinds of validity

- Face validity
- Content validity
- Construct validity
- Consequential validity
- Predictive validity
- Concurrent validity

What do you think each of these means?



Six kinds of validity

- **Face validity:** the extent to which the test 'looks right'
- **Content validity:** how closely the content of the test matches the content of the syllabus it is designed to assess



Six kinds of validity

- **Construct validity:** the extent to which the test measures what the test specification or syllabus says it is intended to measure i.e. the match between teaching and assessment
- **Consequential validity:** the extent to which test outcomes can be relied on to make decisions about learners



Six kinds of validity

- **Predictive validity:** the extent to which test results can be used to predict future behaviour or achievement
- **Concurrent validity:** how closely the test outcomes match outcomes from similar tests designed to test related achievement or performance



Task

What kind(s) of validity do these activities involve?

- A school teacher trying to establish whether her internal school mock tests test the same things as O level in the same subject.
- A test designer checking whether a test can show that one group can do something and another group can't.
- Test designers trying to decide whether to include questions about rivers in a new Geography test.
- A teacher investigating whether past students' exam scores reflect the class of university degree they obtain.
- A group of teachers deciding how to weight different sections of a test that each cover different aspects of their teaching syllabus.



Why is validity important?

- Need to be sure that the test is closely related to the syllabus it is intended to test
- In order that the results / outcomes can be used with confidence
- In order to satisfy all stakeholders of the soundness of the examination and the reliability of the scores (accountability and credibility)



How to check for validity

- Test validation
- A form of evaluation based on evidence collected before, during and/or after the test is taken
- Collection and analysis of data about the candidates and their test scores



Ways of ensuring validity

- Clear test specifications closely matched to teaching and assessment syllabuses
- Use of setting grids when test setting
- Thorough vetting and checking of draft test materials (using checklists)
- Careful review of test outcomes
- Seek feedback from learners, markers and other stakeholders



Reliability – some definitions

- Consistency
- Dependability
- Repeatability
- ‘The extent to which assessment results are an accurate measurement of the candidates’ demonstration of the abilities specified by the assessment criteria’ (QCA)



Reliability – another definition

- ‘The degree to which exam marks are free from errors of measurement and the extent to which they can be depended on for making decisions about the candidate.’

(Cyril Weir, 2005)



Validity and reliability

- Validity and reliability are very closely linked
- Some people consider that reliability is another form of validity
- There may be tensions between validity and reliability
- An unreliable test cannot really be valid



How to check for reliability

- Repeated measurements or observations are required, so that we can compare outcomes from a repeat of the same test on the same subjects
- But, in practice, we can't measure the same learners again and again using the same test (learners are not machines!)



How to check for reliability

- Test lots of learners on one occasion
- Test the same learners on two occasions with the same test
- Test learners on different versions of the test



Statistical tests of reliability

- Correlation between a test and a parallel test
- Correlation between one test taken on two occasions (test-retest)
- Internal consistency reliability (split test in two)
- Cronbach's alpha (multiple split test)



Reliability in assessment

Reliability is required in various aspects of assessment:

- The assessment instrument / test
- Test paper production
- Test administration
- Marking
- Awarding / grading



Factors affecting reliability

- Test length
- Question type(s) used
- Quality of questions
- Quality of mark scheme
- Care taken with test paper production
- Test administration:
 - Instructions to candidates
 - Time available
 - Disruptions



Factors affecting reliability

- **Marker reliability**
 - Standardisation (inter-marker reliability)
 - Mark scheme
 - Quality of marker
- **Awarding reliability**
 - Number of grades
 - Use of statistical information
 - Use of archive material (reliability over time)



Ways of ensuring reliability

- Clear test specifications
- Use of specification grids when test setting
- Thorough vetting and checking of draft test materials (checklists)
- Limited question and option choice
- Effective standardisation and monitoring of markers
- Careful review of test outcomes



Task

Which of the following is likely to give reliable results?

- A computer-marked multiple choice test
- Examiners marking tests in a centralised venue
- Giving examiners a quota of scripts to mark per hour
- A test consisting of short answer questions
- An essay question 'Football is more important than politics. Discuss.'
- Giving examiners freedom to interpret marking criteria
- Requiring examiners to attend a standardisation meeting



What is fairness in assessment?

- Fair = just, equitable, free from bias
- ‘Fairness is concerned with the consequences of testing for individuals, groups or society as a whole. It relates to both the validity of a given test as an index of ability and to the whole testing process in so far as it reflects or contributes to social equity.’



QCA principles of fairness

- Assessment instruments are designed to make sure that all candidates are treated fairly.
- The assessment system is transparent so that:
 - The evidence on which decisions are based is clear
 - It is possible to audit the decision-making process
 - Those making decisions can be held accountable



QCA on Fairness

- In examinations aimed at a wide cohort – like GCSE and GCE – examiners aim to address the full range of candidate attainment, need and background while holding to established standards. To promote equal opportunities, examinations must accordingly:

.....cont'd



QCA on Fairness (cont'd)

- be free from gender, ethnic or other discrimination and stereotyping
- use content, resources and assessment materials that recognise the achievements and contributions of different groups
- permit alternative learning and teaching approaches
- provide an appropriate balance of assessment methods



Bias in assessment

Bias = being unfair to one particular group:

- Gender
- Ethnicity / race
- Age
- (Dis)ability
- Social class / background
- Nationality
- Language
- Other?



How to be fair & avoid bias

- Articulate the construct of the test clearly
- Ensure that the assessment accurately reflects the construct (validity)
- Avoid content that is not relevant to the construct
- Make assessment criteria clear to teachers and learners
- Use a range of contexts, test types, content, response formats

(Gipps & Stobart, 2004)



Task

How might the following affect a test's fairness?

- Choice of font and font size
- Use of simple language
- Inclusion of complicated diagrams and maps
- Timed essays
- Holding a test at 8pm
- Holding a test in a comfortable venue
- Multiple choice format
- An all male or all female examining team



Manageability – a definition

- The extent to which an assessment is practicable, implementable, affordable and repeatable



Manageability for whom?

- School
- Teachers
- Learners
- Markers
- Others?



Manageability in terms of what?

- Resources for development and implementation of test:
 - Physical
 - Intellectual
 - Personnel
 - Technical / computer
- Time
- Cost of development and administration
- Complexity



Factors in manageability and cost

- What makes an assessment more or less manageable and expensive to run?



Factors in manageability and cost

- Availability of necessary expertise
- Ease of administration
- Ease of marking
- Human resources required
- Computer resources required
- Complexity of scheme / question paper / test



Factors in manageability and cost

- Number of options / optional routes
- Amount of question choice
- Length of question paper / test
- Frequency of testing
- Amount of source material required
- Amount of artwork, photographs, etc required
- Copyright costs