Are our assessments really valid?

Using validity paradigms to design and evaluate programmes of assessment

Workshop 0C-6:
16th Ottawa Conference
April 2014

Trudie Roberts
Director of Leeds Institute of Medical Education
University of Leeds Medical School, UK

Katharine Boursicot
Assistant Dean for Assessment
Lee Kong Chian School of Medicine
Nanyang Technological University
Singapore

Richard Fuller
Director of Undergraduate Medical Studies
University of Leeds, UK

How well is your assessment working?

- Is it valid?
- Is it reliable?
- Is it doing what it is supposed to be doing?

- To answer these questions, we have to consider the **characteristics** of assessment instruments

Define the purpose/use of the assessment

Characteristics of assessment instruments: Utility function

$$U = {}_{w_r}R \times {}_{w_v}V \times {}_{w_e}E \times {}_{w_a}A \times {}_{w_c}C$$

- U = Utility
- R = Reliability
- V = Validity
- E = Educational impact
- A = Acceptability
- C = Cost
- W = Weight

Van der Vleuten, CPE & Schurwith, L
2013

Developing modern views of validity

- Overarching unitary concept (Messick, 1989)
  - 'all validity is construct validity'
- Interpretative argument (Kane, 1994)
  - 'test data has little or no intrinsic meaning'
- Standards of Educational and Psychological Measurement (AERA, APA & NCME, 1999)
  - 'Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests'
- Evidence to support the interpretation of assessment data (Downing, 2003)
- A 2 stage process…(Kane 2013)

Stage 1: intended use argument

- What is your assessment intended to achieve?
  - To permit progression/graduation/licensing?
  - To act as an assessment for learning / developmental exercise?
  - To detect unprofessional behaviour?
- How explicitly is this stated (and understood?)
- Does this argument 'hold' across all stages of the assessment process?
  - Intention / design
  - Delivery
  - Analysis
  - Consequences and outcomes (e.g. misusing formative data)

Stage 2: Gathering meaningful evidence for the validity of a test

Does the assessment measure what it is intended to measure? Evidence to be collated:

1. Content
2. Response process
3. Internal structure
4. Relation to other variables
5. Consequences

1. Content

- Work in groups of 2 or 3
- Choose a test or programme of assessment from your own institution with which you are involved
- Take 5 minutes to discuss how you decide on the content of a test and the quality control

1. Content

- Blueprint:
  - Mapping test domains to curriculum objectives
  - Matching item content to domains of blueprint
  - Test specifications (eg numbers of items, hours of testing)

- Quality control of test items
  - Item writing process
  - Review process

2. Response process

- What are your safeguards in relation to getting the student score right?
  - Quality control of
    - scoring (electronic scanning/electronic scoring/markers)
    - combining scores (compensation across different formats?)
    - applying pass-fail rules correctly & accuracy of final scores/ marks/grade
    - score reporting to examinees/faculty

- Familiarising examinees with test format

- Enabling examinees to understand the examination regulations and their scores

Workshop task

- Continue in your groups of 2 or 3
- Use the same test or programme of assessment from your own institution with which you are involved
- Apply the validity framework section 2 (handouts)
  - Questions are asked to help you think about your assessments and collate your evidence
  - If you don't know or can't answer – don't worry!
- You have 15 minutes for this section

3. Internal structure

- Looking at the psychometric properties of your test:
    - Item analysis data
    - Reliability
    - Standard Error of Measurement (SEM)
    - Generalizability studies
    - Item factor analysis
    - Differential Item functioning (DIF)

Workshop task

- Continue in your groups of 2 or 3
- Use the same test or programme of assessment from your own institution with which you are involved
- Apply the validity framework section 3 (handouts)
  - Questions are asked to help you think about your assessments and collate your evidence
  - If you don’t know or can’t answer – don’t worry!
- You have 15 minutes for this section

4. Relationship with other variables

- Correlation with other relevant variables
- Convergent correlations-internal/external: similar tests
- Divergent correlations-internal/external: dissimilar measures
- Generalizability of evidence

Workshop task

- Continue in your groups of 2 or 3
- Use the same test or programme of assessment from your own institution with which you are involved
- Apply the validity framework section 4 (handouts)
  - Questions are asked to help you think about your assessments and collate your evidence
  - If you don’t know or can’t answer – don’t worry!
- You have 15 minutes for this section

5. Consequences

- Do you formally consider and document the consequences of your test?
  - Standard setting method explicitly stated
  - Impact of
    - failing (effect on students/institution)
    - passing (effect on professional standards/ society)
  - Impact on learners and future learning (feedback)
  - Impact on faculty/teaching

Workshop task

- Continue in your groups of 2 or 3
- Use the same test or programme of assessment from your own institution with which you are involved
- Apply the validity framework section 5 (handouts)
  - Questions are asked to help you think about your assessments and collate your evidence
  - If you don't know or can't answer – don't worry!
- You have 15 minutes for this section

Formative or summative?

- Traditional Psychometric models have a critical role in determining validity – but pose threats to interpretation in highly contextualised assessments (best evidence = WBA)
- Intended Use Argument – how is this meaningfully disseminated (Haertel, Brennan 2013)
- What evidence do I need to capture (other than an anticipated low reliability)?
- Learning to Love the subjective
  - Student engagement and behaviours?
  - Response – global assessor judgements and expertise (Crossley 2012)
  - Narratives - more words, less numbers (Govaerts 2014)

Professionalism

- Accepting ‘psychometrically unreliable ‘evidence sources and applying some pragmatism

- Lapses vs. consistently unprofessional behaviour
- Multiple sources, over time = not ‘the professionalism WBA’
- Multiple assessors, realities and judgements (Govaerts 2007, Gingerich, Kogan 2011)
- Do we want to wait to see that someone is ‘reliably’ of major concern?!

Validity: a QA model and tool

Collate and interpret evidence for the validity of a test in a meaningful way, to answer the question:

Does the assessment measure what it is intended to measure?

Evidence to be collated:

1. Content
2. Response process
3. Internal structure
4. Relation to other variables
5. Consequences

+

Thank you for your attention

t.e.roberts@leeds.ac.uk

r.fuller@leeds.ac.uk

katharine.boursicot@ntu.edu.sg