Ensuring Quality and Standard in Assessment

Justin C Konje
Sidra Medical and Research Center
PO Box 26999
Doha, Qatar
Why undertake an assessment?

Assessment
- Are students fit for purpose (GMC)
- Have the appropriate training goals been achieved?
- Should progression in training be allowed

Quality Assurance
- Are assessments fit for purpose
  - Peer/External Examiners
  - Blueprinting
  - Post-exam item analysis
- Enhanced Annual Review (EAR)
How should assessment be undertaken?

Reduce Costs in designing assessments

- Reducing time – supporting process in software
  - Peer/External examiner review
  - Standards Setting
  - Blueprinting
  - Item Analysis
Who is involved in the process?

**People**
- Staff – Academic
- Staff – Administrative
- Staff – Technical
- Students

**Purpose**

**Process**
Result – have the aims been achieved?
Setting of Question Papers
Objective of setting/designing assessment

Blue Print

Physical Assembling

Attributes of Good Test

Detailed Question wise analysis

Appropriate marking scheme

Aim: To have a robust process
Attributes of good of designing a good assessment

- Clarity
- Reliability
- Validity
- Authenticity
- Fairness

The goal – Test items should be really difficult for people who don't understand the subject material, but they should be straightforward for those who do. If an item is difficult because of complicated wording (e.g., double negatives) or vocabulary, you will end up testing language skills rather than ability in the discipline.
Attributes of good question design

Clarity

Reliability

Validity

Authenticity

Fairness

“Nothing in the content or structure of [a test] item should prevent an informed student from responding correctly.” Gronlund (1998)

Think about points such as:

- unclear test instructions,
- confusing and ambiguous terminology,
- being overly verbose,
- using complicated vocabulary,
- difficult or poor sentence structure,
- unnecessary and distracting detail.
Version B

Compare the advances in UK public health policy pre- and post-Second World War?
Attributes of good question design

• Does the question allow markers to grade it consistently and reproducibly and does it allow markers to discriminate between different levels of performance?
• Depends on the quality of the marking guidance and clarity of the assessment criteria.
• The likelihood of eliciting an accurate measure of a student’s ability will be increased when students are provided with a variety of ways to demonstrate their knowledge and skills. i.e Triangulate Question Formats
Attributes of good question design

- Clarity
- Reliability
- Validity
- Authenticity
- Fairness

A valid examination question measures achievement of the intended learning outcomes of the module/unit module (not just what is easy to measure).

The form of the examination question may also be of importance in ensuring validity.

For example, Short Vs Essay Type Question

- Short Answers: Good way of assessing greater breadth of material covered in a course and tend to focus on testing attainment and application of knowledge.
- Essay style questions: Allow a more in depth exploration of subject material and require a candidate to build and structure an argument.
Attributes of good question design

Clarity | Reliability | Validity | Authenticity | Fairness

Authenticity is the need to match the style and approach of question setting to the reality of practice. This may be very important when considering the testing of “procedural knowledge“ or “functioning knowledge“.

For Example:
- EMQ on the step-wise approach to the investigation/Rx of a woman with suspected DVT/PE

Variations:
- What initial investigation on suspicion of PE
- What if the CXR is abnormal – which test to perform?
Attributes of good question design

You need to give students a fair chance to demonstrate what they know and can do and to be able to succeed in examinations. Fairness can be facilitated by:

• Being very clear about expectations in student performance,
• Providing examples of past examination papers,
• Opportunities for students to practice and gain → Mock Exams
• Transparency in the processes to mark and grade their work.
What kinds of knowledge and skills can be tested in examinations

“Exam questions should test a range of knowledge and skills at Masters level. They should test and reward critical appreciation and the ability to apply what has been learnt rather than the passive reproduction of memorised facts.”

Assessment Code of Practice, (2012)

kinds of knowledge that can be tested
kinds of intellectual skills
kinds of transferable skills
kinds of attitude

e.g. Knowledge domains
     e.g. Analysing, Evaluating
     e.g. Writing skills, Time use
     e.g. Ethics, equality
5.1. The kinds of knowledge that can be tested – knowledge domains

- Factual Knowledge
  *Terminology, facts, figures*

- Conceptual Knowledge
  *Classification, Principles, Theories, Structures, Frameworks*

- Procedural Knowledge
  *Algorithms, Techniques and Methods and Knowing when and how to use them*

- Metacognitive Knowledge
  *Strategy, Overview, Self Knowledge, Knowing how you know*
Objective of setting/designing assessment

- Blue Print
- Physical Assembling
- Attributes of Good Test
- Detailed Question wise analysis
- Appropriate marking scheme

Aim: To have a robust process
Concept of Blue printing

- The blueprint – word borrowed from architecture.

- *It indicates that a process of assessment needs to be conducted according to a replicable plan,*

- the use of the blueprint will ensure that, the test has been developed and mapped carefully against the educational objectives of the course, ensuring fair representation of objectives.
Paper setting and blue Printing

1. Define Objective: the purpose for which blue print is to be made
2. Paper to be set for internal assessment or Final examination or an entrance examination
3. Syllabus or course content according to the level of the student
4. Time allotted and number of questions which can be done within stipulated period or time
5. Decide on the design of the paper
6. Types of questions in a question paper: only MCQs or mixed
7. Questions to be framed keeping in mind the Taxonomy of educational objectives & table of specifications.
8. If the paper has mixed percentage: distribution of questions & distribution of marks: Marks to be distributed according to the question type, expected weightage on the content.
Objective of setting/designing assessment

Aim: To have a robust process
What is the Passing Score?

- Essential component of high stakes exams
- Reaffirm standards
- Their purpose is to ensure that
  - qualified candidates pass
  - unqualified candidates do not pass
- How much is enough?
- Is 50% the passing score on this exam?
The basis for passing scores

• Arbitrary judgment unavoidable
• Reflect consensus of experts on reasonable expectations for evidence of competence
• Imposing discrete categories on a continuum
• Set to serve the interests of public and profession
• Process should be as open as possible
• Based on as much relevant data as possible
• Rationale presented as clearly as possible
Process of Setting Passing Scores

• Unreasonable to expect 100% correct
• Possible to construct tests with predetermined passing scores
• Possible to adjust passing scores to achieve an acceptable pass rate
• Possible to estimate a minimum passing score by combining estimates of the importance of individual test items
Passing Score Level

- Determined by the situation and purpose
- Provide society with enough sufficiently competent practitioners
- Raising the passing score increases the average competence of those who pass but decreases their number
- Proportions passing should remain constant
- The more relevant and demanding the requirements for writing the test, the fewer are expected to fail
- If more than a small proportion of successful candidates fail the exam, its validity may be subject to serious challenge.
However....

- Fails to take into consideration variables:
  - Content of examination
  - Examiners
  - Candidates
  - Teaching
  - Other factors
Imprecision in Determining a Pass mark
NOT ENOUGH!

Content Standards

Performance Standard

Test Specification

Test

?
Standard Setting For Post Graduate Examinations

• The **standard** does not change but the **pass mark** does
• No relation to the relative strength of the cohort sitting the examination
• The aim of Standard Setting is to improve the **fairness, reliability and validity** of the examination process. It is not designed to restrict the number of successful candidates.
Important Factors in Setting a Standard

- Subject Matter Experts
- A Representative Committee
- Ample Time
- Actual Test Items (Questions)
- Definition of the Level of Proficiency
  Distinguishing Qualified and Unqualified Candidates
Criteria for defensibility

A standard setting method should …

• produce appropriate classification information
• be sensitive to candidate performance
• be sensitive to instruction
• be statistically sound
• identify the “true” standard
• be easy to implement and compute
• be credible and easily interpretable by lay people
Methods of Standard Setting

1. **Test-centred methods**
   Standards derived from hypothetical decisions based on test content before the test is answered.

2. **Examinee-centred methods**
   Standards derived from reviewing examinees’ performance before deciding cut-off score.

3. **Compromise methods**
   Provide flexibility for adjusting the standard based on the examinees’ performance on the test.
Appropriate marking scheme - how reproducible/fair is the assessment?

Assessment criteria describes the extent to which students have achieved the specified learning outcomes.

Two Different Approaches to Marking

Assigning grades fairly and robustly is a demanding occupation for all teachers and a range of approaches is applied to help do this reliably and consistently.

Two very different methods are often used simultaneously and symbiotically – norm referencing and criteria referencing.
How should Marks be Interpreted?

- Norm-Referenced
- Criterion-Referenced
  - Standards-Based
Norm Referencing

Norm referencing is all about comparison

Ultimate form of norm-referenced assessment is when we attempt to fit our marking profile for a cohort of students to the “bell-shaped” curve

This pattern of achievement anticipates that a few students will fail and a similarly few students will get distinctions whilst the majority will gain marks that cluster and peak in the middle mark range.
Criterion referenced grading

Specifies a standard through the description of clear criteria and anybody who achieves the level or standard described gains the marks – so everybody in the cohort could potentially get an “A”
### Norm – & Criterion-References Standards

<table>
<thead>
<tr>
<th>NORM-REFERENCED</th>
<th>CRITERION-REFERENCED</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Relative</td>
<td>• Absolute</td>
</tr>
<tr>
<td>• Based on peer-performance</td>
<td>• Not related to peer performance</td>
</tr>
<tr>
<td>• Varies with each group</td>
<td>• Standard set prior to exam</td>
</tr>
<tr>
<td>• Cut-off point not related to</td>
<td>• Referenced to a defined</td>
</tr>
<tr>
<td>competence</td>
<td>level of performance</td>
</tr>
</tbody>
</table>
More than 3 dozen methods

Some of the better known methods include:

- Nedelsky
- Angoff
- Bookmark
- Ebel
- Jaeger
- IRT (item response theory) methods
The Angoff Method is:

- the most commonly used method
- convenient to use
- well-researched
- easily explained
- easily customized
- applicable to several response formats
The Angoff Method

- Judges assign probabilities that a hypothetical minimally competent borderline candidate (an average good trainee) will be able to answer each item correctly.
- For each judge, probabilities are summed to get a minimum performance level (MPL).
- MPLs are averaged to get a final passing score.
Minimally Competent

- The effectiveness of the Angoff method rests on the judges’ ability to accurately conceptualize a “minimally competent, borderline candidate.”
- Repeated references to a formal summary of the behaviours and performance indicators is required
- Judge training and calibration are essential
Angoff Calculations

<table>
<thead>
<tr>
<th>Item</th>
<th>Judge 1</th>
<th>Judge 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>0.85</td>
</tr>
<tr>
<td>2</td>
<td>0.65</td>
<td>0.50</td>
</tr>
<tr>
<td>3</td>
<td>0.80</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>0.45</td>
<td>0.50</td>
</tr>
<tr>
<td>5</td>
<td>0.30</td>
<td>0.40</td>
</tr>
<tr>
<td>MPL</td>
<td>3.2</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Passing score for this test is 3.1 items correct out of 5.
Angoff with Iteration

- Most commonly used modification.
- “Angoff-ing” is done a number of times.
- Time between rounds is used for discussion among judges.
- Intent is to reduce variability among judges on item estimates.
Subject Matter Experts define the minimally competent (borderline) candidate in terms of knowledge.

They evaluate every item in the test and estimate this candidate’s chances of answering correctly.

The mean estimate across all experts and all items determines the passing score.
Advantages of the Angoff Method

- A relatively straightforward process
- No data necessary
- Has held up in court
Disadvantages of the Angoff Method

- Must look at every item on test(s)
  - Time and cost
  - Fatigue, inattention, “rushing”
- Difficulty of accurately estimating probabilities
Other Methods

- Borderline group method – used for the oral and now Part 3 MRCOG Examination
- Contrasting group method
- Regression based standard method


The examiner’s role in standard setting

- Uses the examiner’s clinical expertise to judge the candidate’s performance

- Examiner allocates a **global judgement** based on the candidate’s performance at that station

- Remember the level of the examination

Pass  
Borderline  
Fail 
Borderline Group Method

Checklist

1. Η σημερινή σημείωση σφηνή είναι σωστή; ✔
2. Κάθε σκηνή διατηρεί θέλομενο θάρσος; ✔
3. Λαθαία κάθε δίδυμο διαλείμματα; ✔
4. Κεφαλή δεν είναι σωστή; ✔
5. Ψηφίστηκε δίδυμο; ✔
6. Η σημείωση του αποτελέσματος είναι ακριβής; ✔
7. Η σημείωση του αποτελέσματος είναι ακριβής; ✔

TOTAL
Pass, Fail, Borderline

Test score distribution

Borderline score distribution

Passing score
Contrasting groups method

Test score distribution

Checklist

1. Ησ σηφσ σφνησσ σφησ σφσ σφ
   ✔
2. Κσκσ σκσµσιθοπθλ θλθµ θ θ κλ
   ✔
3. Λαλκα κδµ δδκκ δλκλ δλλδ
   ✔
4. Κεψω δδ ε ρ ρρτ τµκ
   ✔
5. Αφφακ δδ
   ✔
6. Ηπκλ ιι σκρ σκα ακ εκλ. αλδ
   ✔
7. Ηδηηδδη σησ αηηακκ ασ
   ✔

TOTAL

Pass, Fail, Borderline: P/B/F

Passing score
Regression based standard

Checklist

1. Ησ σηϕσ σφνησσ σφησ σφσ  σφ
2. Κσκσ σκσµσιθοπθλ θλθµ θ θ  θκλ
3. Λαλκα κδµ δδκκ δλκλ δλλδ
4. Κεψω δδ ε ρ ρµτ τµκ
5. Αφρε δδ
6. Ηπκλ,σ σκφ σλσ σκα ακ εκλ. αλδ
7. Ηδηηδδη σησ αηηακκ ασ

TOTAL

Overall rating  1 2  3  4  5

Σ

X = passing score
Performance-based standard setting

- Disadvantages
  - Requires large cohort of candidates to achieve enough numbers in the ‘borderline’ group
  - Passing score not known in advance
  - Judgments not independent of checklist scoring
  - Requires expert processing of marks immediately after the exam
    - Checking of results
    - Delay in producing results
THANK YOU