Understanding Standard Setting

Daniel Zahra
Jo Cockerill

daniel.zahra@plymouth.ac.uk
• **Who** standard sets?

• **What** do you standard set?

• **How** do you standard set?

• **Why** do you standard set?
The purpose of standard setting is to determine the pass mark for a test.

Friedman Ben-David (2000)

Meaning varies with assessment aims, e.g.

- Progression
- Competence
- Accreditation
Aims

• **Overview** of standard setting methods

• **Advantages and Disadvantages**

• Consideration of **application**

• Discussion of **applicability**
Reference Points

**Norm Referenced (Relative)**
Relative to the performance of the group

**Criterion Referenced (Absolute)**
Relative to an external standard

**Compromise**
Combination of norm- and criterion-reference
Norm-Referenced

Based on group performance.

Does not evaluate competence with respect to external benchmarks/criteria.

For example, the pass-mark may be set at the mean score, or only the top 10% awarded a distinction.
Criterion-Referenced

Independent of the group taking the test.

Standard is based on predetermined level of competency.

For example, the pass-mark may be set at 60% ahead of the exam.
Compromise Methods

These combine both norm- and criterion-referenced standards.

For example, we might set a standard at 40%, but adjust it by mean difference across tests.
Methods

Norm Referenced
Set Proportions, SD from Mean

Criterion Referenced
Fixed standard, Angoff, Ebel

Compromise
Hofstee, Hofstee-Angoff, Contrasting Groups
Borderline Groups, Borderline Regression
Driving Theory Test

Different methods of standard setting an assessment...
If you were to set a single, unchanging pass-mark for the test, what would it be?
Fixed Standard
Fixed Standard
Fixed Standard
Expert judge provides estimates of performance
  • Proportion correct
  • Yes-No

Borderlines / minimally competent / similar

The standard is then the average item estimate.
For each item, imagine a minimally competent candidate and ask yourself whether they would answer the item correctly.

Record you decision next to each item with a Yes or No.
Discuss your responses as a group and review individual judgements.
Averaging across your ratings, we set a pass-mark of...
Requires a min/max pass mark
Requires a min/max failure rate

The pass mark is located where the student data intersects these parameters on a cumulative plot.
Cumulative Frequency Curve for Stage 2 Scores
What would you accept as the:
• Minimum Pass-mark
• Maximum Pass-mark
• Minimum Failure Rate
• Maximum Failure Rate
What happens if you want to add grade boundaries?

- Fixed pass-mark (e.g. >=75% = Distinction)
- Error/Variability (e.g. Cut-score +/- SEM)
- Set Proportions (e.g. Top 5% = Excellent)
- Relative boundaries (e.g. Mean +/- Value)
What happens if you also have global grade* judgements being made alongside numerical scores?

“This student scored 68%, and I thought their overall performance was Satisfactory”

*Global-grades which provide an overall, more subjective/ecdlectic/holistic judgement of performance; not necessarily a grade derived from the score (e.g. 75% = A) or otherwise tied to the criteria used to derive the scores.
Borderline Groups uses the average of the B graded scores. Contrasting Groups sets the pass-mark between U and LS grades.
Borderline Regression uses the scores and grades to derive a linear model, and sets the pass-mark at the intersection of the borderline group.
There is no gold standard.
The method must be:

- Appropriate
- Feasible
- Credible
- Acceptable to stakeholders
- Evidence-based
- Defensible (academically and legally)
Recap

Norm-Referenced

- Provides standardised pass/fail rate
- Easy to implement
- Does not adjust for ability
- Might not be acceptable to all stakeholders
Recap

Criterion-Referenced

- Focus on individual items
- Defensible for high stakes
- Borderline can be difficult to define
- Time-consuming
- Where do the numbers come from?
Recap

Compromise Methods

• Suitable for overall pass/fail
• Evidence based
• Simple standard setting
• Can ‘miss the mark’, prone to outliers
• Not first choice for high-stakes
Conclusions

There is no perfect method.

A wide range of methods exist.

Methods must be fit for purpose.

Can be a question of policy.
Conclusions

Choice depends on:
  • Credibility
  • Available resources
  • High stakes level of exam

Method is important, process is critical
  • Suitable judges
  • Due diligence applied
  • Defensible rationale
daniel.zahra@plymouth.ac.uk
jo.cockerill@plymouth.ac.uk