SETTING STANDARDS

Doan Thi Thu Hoa, MD, MHPE
Faculty of Medicine, UMP
09.2023

Objectives

- 1. Differentiate types of standards (relative/absolute standards)
- 2. Analyze process to implement Angoff method for setting standard in a MCQ test
- 3. Analyze process to implement borderline group method for setting standard in an OSCE

Standards

- The standard is an artificial, but necessary, dichotomy imposed on the continuous variable of competence in order to make a decision pertaining to the competence of an individual in the given field in which a test is taken.

(AMEE guide 18)

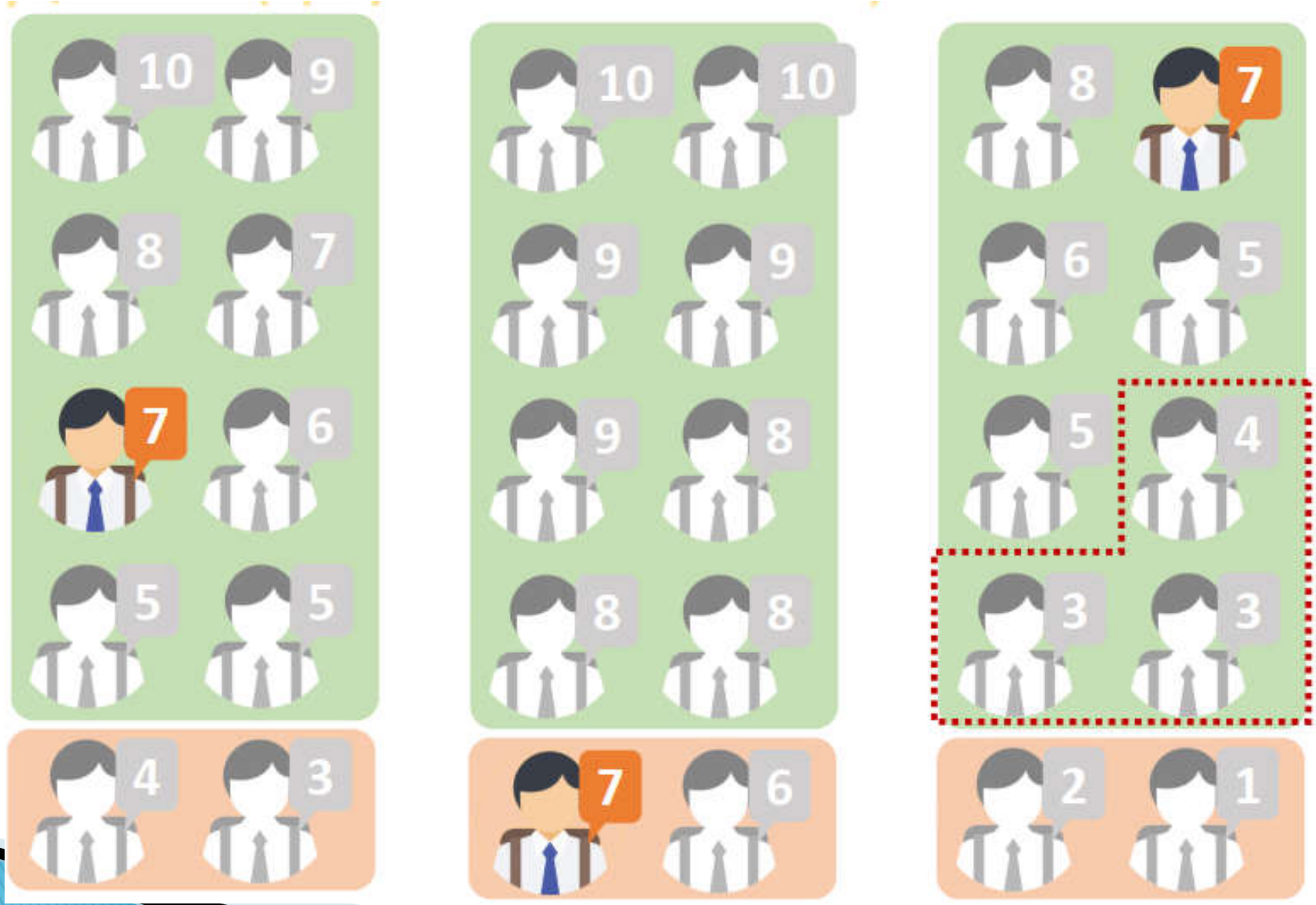
- A pass/fail decision on an examination enables the separation of competent and incompetent candidates

Types of Standards

- **Relative standards/ norm-referenced standards:**
  - Based on a comparison among the performances of examinees
  - A set proportion of candidates fails regardless of how well they perform
  - e.g. the top 80% pass
- **Absolute standards/criterion-referenced standards:**
  - Based on how much the examinees know
  - Candidates pass or fail depending on whether they meet specified criteria
  - e.g. examinees must correctly answer 70% of the questions

norm-referenced standards

Top 80% pass



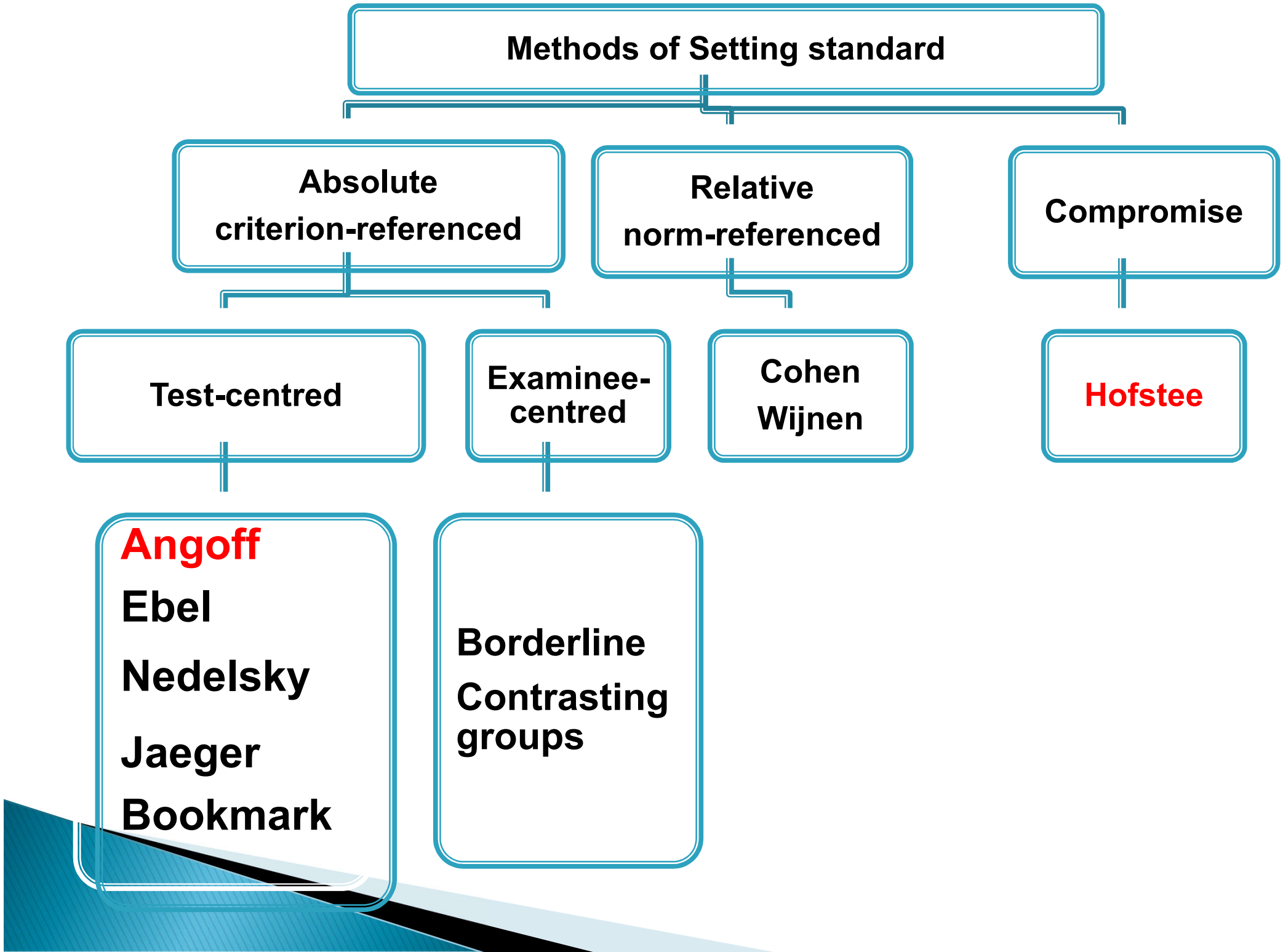
criterion-referenced standards

70% correct answer

Weaknesses of relative/norm-referenced standards

- Standards are not content related
- Fixed number of candidates may fail each year
- Examinees' ability may influence the standard
- Standard is not known in advance
- Diagnostic feedback relative to performance is unclear

AMEE guide 18


Methods of Setting standard
Absolute criterion-referenced
Relative norm-referenced
Compromise
Test-centred
Examinee-centred
Cohen
Wijnen
Hofstee
Angoff
Ebel
Nedelsky
Jaeger
Bookmark
Borderline
Contrasting groups

Criterion-referenced standards

- **Test-centered standards** are those derived from hypothetical decisions based on the test content.
- **Examinee-centered standards** are those derived from reviewing the performance of examinees or a similar group prior to making judgments about what constitutes borderline performance between competence and incompetence.

Minimally Competent Candidate/ Borderline candidate

- Candidate **possessing the minimum level of knowledge and skills necessary to perform at a registration/licensure level.**
- This candidate performs at a level **"on the borderline" between acceptable and unacceptable performance.**
- Assigning a candidate to a category of "borderline" in assessment implies that the **assessors are unsure if the candidate is neither clearly satisfactory nor clearly unsatisfactory**
- It is essential that each judge arrive at a clear and specific definition of the minimally competent candidate.

- Judges often start off with different interpretations of this construct. The terms must be thoroughly discussed by the group of judges with the help of a facilitator. **A common understanding must be arrived at, before independent judgments are made. Failure to do so would likely result in a divergent set of cut-off scores**
- **(AMEE guide 37)**

- Borderline candidates are more likely to score well on **topics that easier to comprehend**, and/or **repeatedly taught or asked.**
- This means that certain topics which are considered by curriculum developers and assessors to be more integral to the core curriculum, are often taught with subjectively greater importance and practically with greater resources of time, and personnel, are more likely to be conducive to a satisfactory performance by a group of borderline candidates

The role of judges in setting standards

- The judges are the key elements in the standard setting process
- The process of judgment is necessarily a subjective one with inherent variability.
- → selection of judges. Criteria for selection include expertise in subject matter, familiarity with similar candidate groups, skill in conceptualizing and skill in self-monitoring.

The role of judges in setting standards

- → adequate training in the process.
  one major difficulty judges have is to understand and keep in mind the construct of "borderline candidate" or "minimally competent candidate".
- → Systematic standard-setting process, which provides ample opportunity for discussion and deliberations among panellists.

Selection of judges/panellists

Panelists should be

- **1. Experts in related field of examination**
- **2. Familiar with examination methods**
- **3. Good problem solvers**
- **4. Familiar with level of candidates**
- **5. Interested in education (teachers)**

Selection of judges/panellists

- **Subject Matter Experts (SMEs)**
- They need to represent a **mix of attributes** (educators-academics vs practitioners, age, gender, seniority…)
- Successful standard setting meetings can be conducted with **as few as 4 panellists or as many as 20.**

Test-centred Methods

- The Nedelsky method (1954)
- The Jagger method (1982)
- The Angoff method (1971)
- The Ebel method (1972)
- Bookmark (1996)

Nedelsky Method

Item	Options					Remaining	Probability
1	~~A~~	B	~~C~~	~~D~~	E	2	.50
2	A	~~B~~	C	D	E	4	.25
3	~~A~~	~~B~~	~~C~~	D	~~E~~	1	1.00
4	A	B	C	~~D~~	E	4	.25
5	A	~~B~~	~~C~~	D	E	3	.33
6	A	~~B~~	C	~~D~~		2	.50
7	A	B	C	D		4	.25
8	A	B	~~C~~	~~D~~		2	.50
9	~~A~~	B	C	D		3	.33
10	~~A~~	B	~~C~~	D		2	.50
Recommended Cut Score for This Rater:							4.41

Figure 2. An example of the Nedelsky Method for 1 Rater and 10 Items

Jagger Method

Should Every Beginning Practitioner Answer This Item Correctly?					
Item	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
1	Y	Y	Y	Y	Y
2	Y	Y	N	Y	N
3	Y	N	N	Y	N
4	N	N	Y	N	N
5	N	Y	Y	Y	Y
6	Y	Y	N	N	Y
7	N	Y	N	Y	Y
8	Y	Y	Y	Y	N
9	Y	N	Y	Y	N
10	N	N	N	N	Y
Total	6	6	5	7	5
Cut Score					5.8

Figure 5. An Example of the Jaeger Method for 5 Judges and 10 Items

Angoff method

There are five steps involved:

- 1. Select the raters/judges.
- 2. Take the assessment.
- 3. Rate the items.
- 4. Review the ratings (Record ratings, discuss, and change)
- 5. Determine the cut score. Calculate the passing score

Angoff method

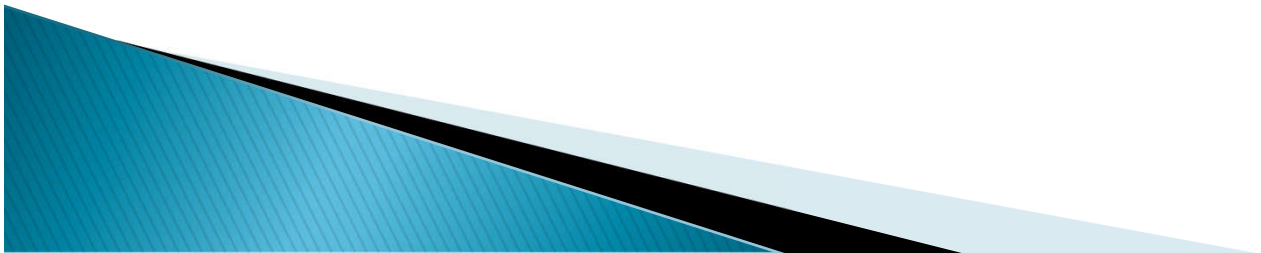
- 1. Select judges (subject matter experts SME)
- Number of judges:
  - Specify the appropriate number (minimum 6-8 for high stakes testing)
  - Ideally: at least 10

Angoff method

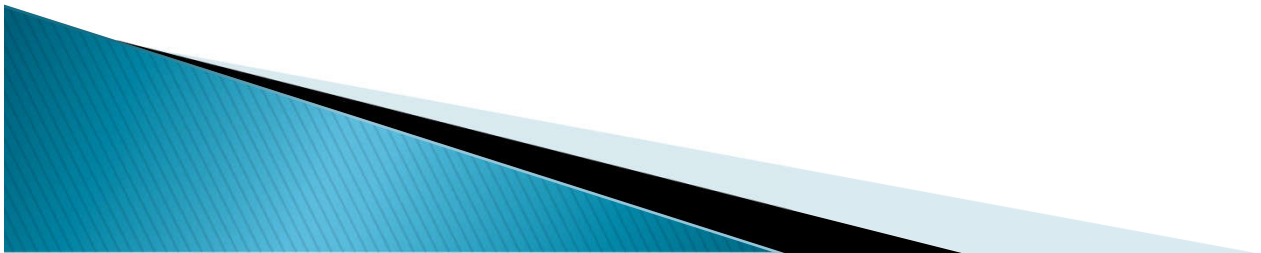
- **1. Select judges (subject matter experts SME)**
- **Discuss**
  - Purpose of the test
  - Nature of the examinees
  - What constitutes adequate/inadequate knowledge
  - The characteristics of Minimally Competent Candidate/Borderline candidate

- **2. Take the assessment.**

- **3. Rate the items**
- **Read the item**
- **Judges estimate independently percentage of minimally competent candidates would answer correctly for each question on the test.**
- **Estimated value ranges from 25% - 95% with increments of 5.**
- **Always keep in mind the concept of borderline candidates**
- **(Of 100 Minimally Competent Candidates, How many will answer this item correctly?)**

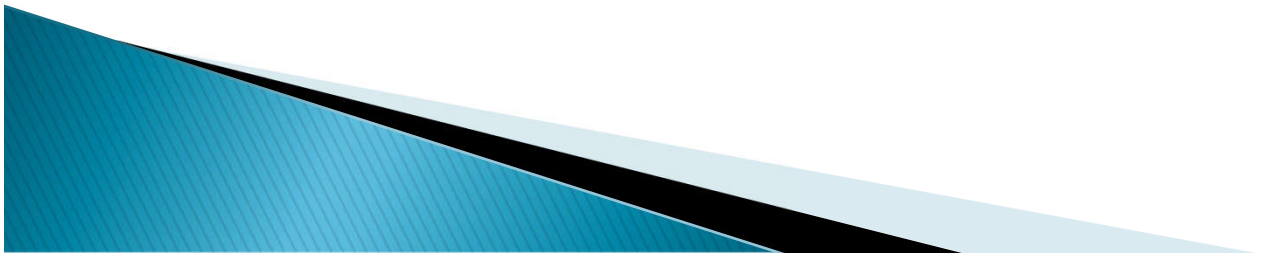


Item	rater 1	rater 2	rater 3	rater 4	rater 5	rater 6	**Average Percentage Correct**	**Standard deviation**
1	50	90	80	60	50	50	63.3	17.511901
2	50	70	60	50	40	40	51.7	11.690452
3	50	60	40	40	40	40	45.0	8.3666003
4	70	70	80	70	60	60	68.3	7.5277265
5	50	90	70	70	40	50	61.7	18.348479
6	50	90	80	60	50	50	63.3	17.511901
….								
….								
….								
Average								



- **4. Review the ratings, discuss and change**
- **The ratings for any single item should be in agreement**
- **By agreement it is meant that the ratings for an item must all be within a certain percentage range (e.g., a 30%-range or standard deviation < 10).**
- **If the range of the ratings is greater than the specified range, the judges providing the extreme ratings are asked to explain why they rated the item in that fashion. The other judges should explain why they rated the item as they did.**

Item	rater 1	rater 2	rater 3	rater 4	rater 5	rater 6	**Average Percentage Correct**	**Standard deviation**
1	50	70	70	60	50	50	58.3	9.8319208
2	50	70	60	50	50	50	55.0	8.3666003
3	50	60	40	40	40	40	45.0	8.3666003
4	70	70	80	70	60	60	68.3	7.5277265
5	50	60	70	70	50	50	58.3	9.8319208
6	50	70	70	60	50	50	58.3	9.8319208
….								
….								
….								
Average	53.5	63.9	62.9	61.9	50.3	50.6	**57.2**	



- **5. Calculate the passing score**

the average rating is calculated for each item and then for the total examination.

average percentage correct

Modified Angoff

- A modification adds to the process by the provision of data about difficulty of the items, based on actual performance data.
- Data are presented to the judges before the final iteration.

Examinee-centred methods

- Borderline group method (1982)
- Borderline regression method (2006)
- Contrasting groups method (1982)

Borderline group method

- Judges are requested to judge a group of individuals as borderline candidates, based on their previous experience or some procedure other than the test itself.
- The scores on the test of these candidates are arranged in rank order and the **median (or mean) score for the borderline group is taken as the cut-off score**.

Borderline group method

- One modification to this method that has been used with OSCEs is to use the judgments gathered during the OSCE administration.
- Examiners mark checklist, and also rate the candidate's performance in global rating such as ‘pass’, ‘borderline’ or ‘fail’
- **Median (or mean) score for the borderline group is taken as the cut-off score**
- **The scores of the borderline group should cluster** together to produce a reasonable standard. If the scores are spread, the method may not be applicable.

Checklist

1st item	✔
2nd item	✔
3rd item	
4th item	✔
5th item	
6th item	✔
TOTAL	Σ

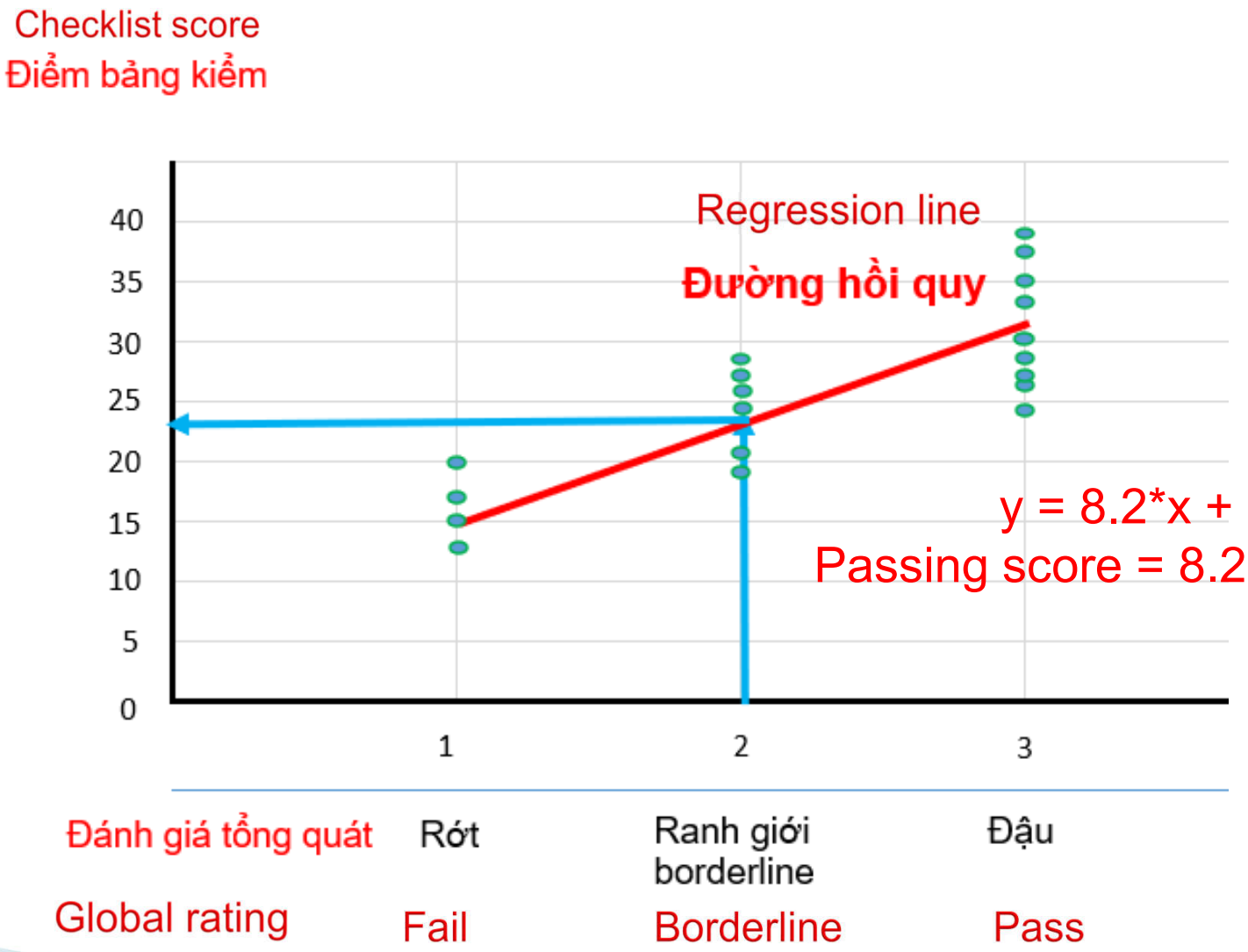
Pass, Fail, Borderline

Borderline regression method

- A regression approach was studied by Kramer (2003), Wood et al. (2006).
- Using the entire range of OSCE scores can be particularly useful if only a small number of examinees have participated.
- In this modification, the checklist score is the dependent variable; the rating is the independent variable.
- The goal of the regression analysis is to **predict the checklist score of the examinees classified as "borderline" for the station**.

Borderline regression method

- the cut-score is calculated using a linear regression equation between checklist score and global score.


Checklist score
Điểm bảng kiểm
Regression line
Đường hồi quy
40
35
30
25
20
15
10
5
0
1
2
3
y = 8.2*x + 7.6
Passing score = 8.2*2+7.6 = 24
Đánh giá tổng quát
Rớt
Ranh giới borderline
Đậu
Global rating
Fail
Borderline
Pass

Selecting a standard setting method

- The method should permit judgments that are based on information; processes that permit expert judgment in light of performance data are preferable.
- The method chosen should be closely aligned with the goal of assessment.
- The method should require thoughtful effort of those participating in the process, and it should be based on research.
- Finally, the method should be easy to explain to participants, and easy to implement.

- There is no perfect standard setting method

No gold standard

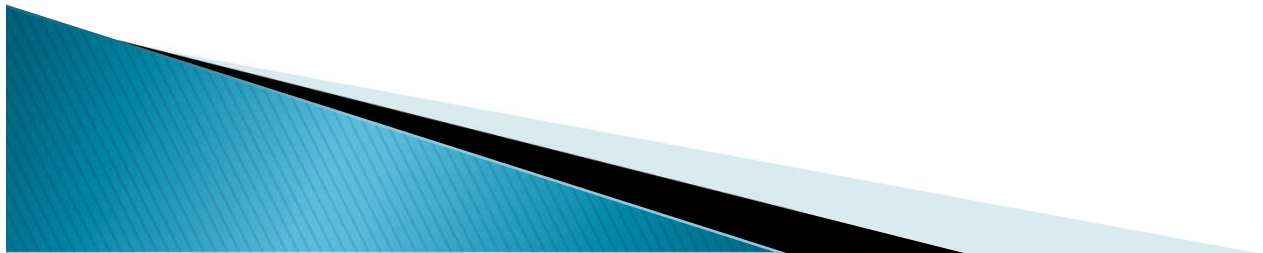
- Make a decision based on the most important criteria for a particular circumstance

Practical implications

- Choice of standard setting methods depends on:
  - Credibility
  - Resources available
  - High stakes level of exam

Method is important, process is critical

- Suitable judges
- Due diligence applied
- Defensible rationale



References

1. Bandaranayake, R. C. (2008). Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. *Medical Teacher*, *30*(9-10), 836-845.
2. Ben-David, M.F. (2000). AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher*, *22*(2), 120-130.
3. Katharine Boursicot, Trudie Roberts. Principles of Standard Setting, pdf presentation
4. McKinley, D. W., & Norcini, J. J. (2014). How to set standards on performance-based examinations: AMEE Guide No. 85. *Medical teacher*, *36*(2), 97-110.
5. Norcini, J.J. (2013). Setting standards. *A practical guide for medical teacher* (4thed.) Elservier.
6. Norcini, J.J. Setting Standards, pdf presentation
7. The Angoff method of standard setting. (2014). https://www.bcrsp.ca/sites/default/files/documents/Angoff%20Method%20Article.pdf


THANK YOU
WISH YOU SUCCESS