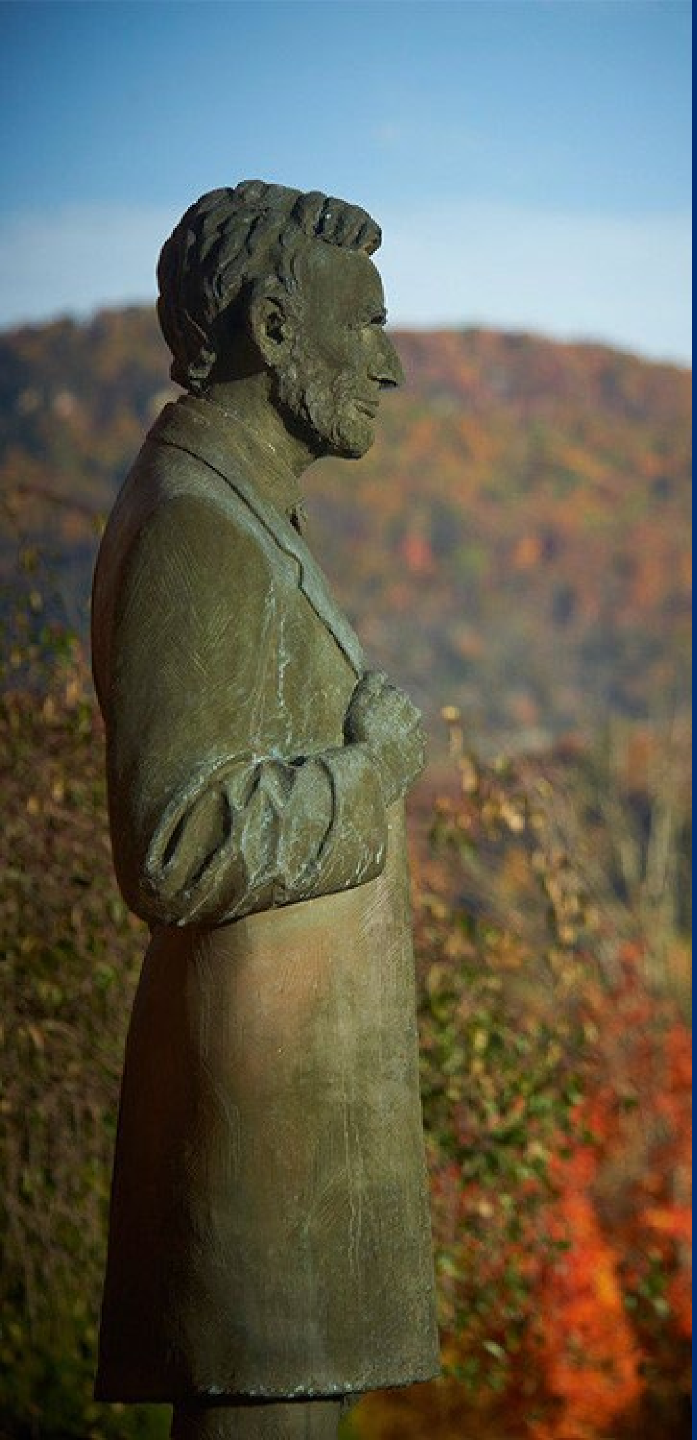


Correct or Incorrect? OSCE Standard Setting and "Grading" Methodologies Utilized in Health Professions Education

IUPUI Assessment Institute
October 11, 2022

Michael Rudolph, Ph.D.
Jill Augustine, Pharm.D., Ph.D., MPH
Justine Gortney, Pharm.D, BCPS

LMU
LINCOLN MEMORIAL UNIVERSITY

Presenters

Mike Rudolph, PhD
Assistant Dean of Academic Affairs and Assistant Professor
Lincoln Memorial University School of Medical Sciences
michael.rudolph@lmunet.edu

Jill Augustine, PharmD, PhD, MPH
Director of Assessment and Assistant Professor
Mercer University College of Pharmacy
Augustine_jm@mercer.edu

Justine Gortney, PharmD, BCPS
Director of Assessment, Division of Pharmacy and Associate Professor
Eugene Applebaum College of Pharmacy and Health Sciences
Wayne State University
justine.gortney@wayne.edu

Learning Objectives

At the completion of this activity, participants will be able to:

1. Differentiate between examinee- and test-centered standard setting methods

2. Review current performance-based assessment practices and develop improvement ideas for standard setting

3. Discuss best practices for standardized grading among a variety of raters

Audience Response Question

How do you currently use OSCEs, if at all, at your school? Select the most appropriate response.

A. As high-stakes assessments ONLY
B. As formative assessments ONLY
C. As high-stakes and formative assessments
D. Do not use OSCEs, but have plans for future use
E. Do not use OSCEs and have no plans for future use.

Google form: https://forms.office.com/r/masLeR7j85

Cut Scores

- **Cut scores** are *selected points on the score scale of an assessment that are used to determine whether a given score is sufficient for some purpose* [1]
- Cut scores are needed when the results of an assessment are used to categorize students in order to make a decision
  - E.g., competent or not competent and progress or remediate

Standard Setting

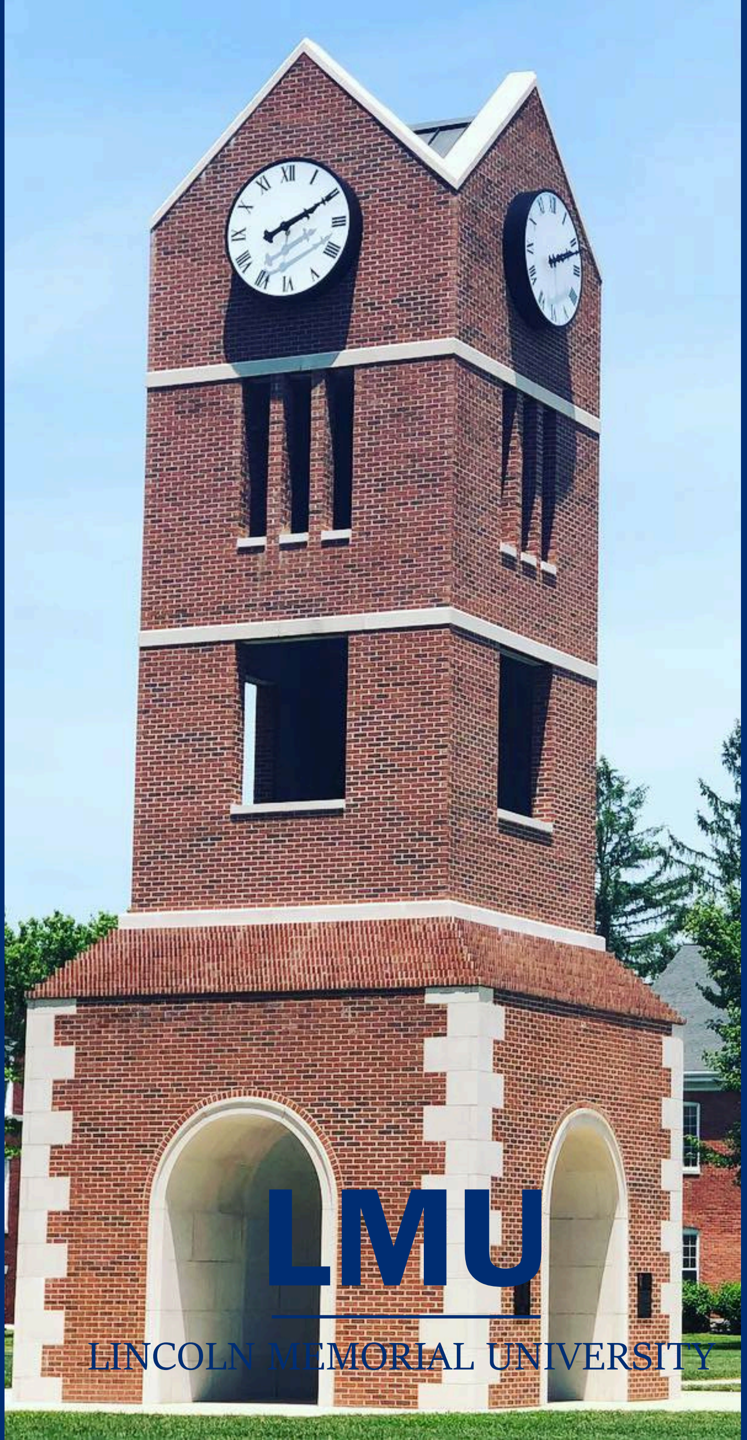
- **Standard setting** is the *process of establishing cut scores on an examination*[2]
- Numerous approaches to standard setting, several of which will be discussed in this session

Why is Having the Right Cut Score Important?

- Making sure the cut score(s) is/are appropriate is an important aspect of validity for the interpretation and use of the test scores[2]
- Cut scores that are:
  - Too low -> passing students who are not competent
  - Too high -> failing students who are competent
- The higher the stakes, the more important it is to correctly categorize students[3,4]

General Approaches to Standard Setting

1. **Norm-referenced standard**: compare performance of student to one or more groups of students, with a fixed number or percent of students automatically passing or failing
   - E.g. bottom 10% will automatically remediate



General Approaches to Standard Setting

2. **Fixed or absolute standard**: judge student performance against a fixed score representing a conceptual definition of competence[5]
    a. **Grade-based**: established using traditional letter grade without consideration for actual ability of examinees or the assessment
    b. **Test-centered**: use judges to review exam items/tasks to estimate the likelihood of borderline students passing
    c. **Examinee-centered**: use judges to review actual student performance on items/tasks to determine if desired level of competence was attained

Importance of Formal Standard-setting

- Norm-referenced or grade-based standards should be avoided with high-stakes assessments due to lack of sensitivity to:
  - Ability level of examinees
  - Difficulty (or easiness) of exam
- Formal standard setting methods base the cut score on the perceived difficulty of the exam items/tasks (test-centered) or actual performance of borderline students (examinee-centered)

Test-centered standard setting

- Also known as criterion-referenced
- Cut-off scores based on expected competence of student on included content
- Advantages
  - Involves experts for judgement
  - Preferred for competency-based assessments
  - Students pass/fail based on expected competence
- Disadvantages
  - Resource intensive
  - Dependent on expert judgement which may be subjective

Test-centered standard setting [2,6]

Method Characteristics	Angoff	Modified Angoff (2 options)	Ebel
Use for clinical-type assessments	• Expert reviews item/tasks and estimates performance of borderline students • Asks "what percentage of borderline candidates would answer this item correctly?" • Mean of experts' score is added and divided by the total number of items to get a cut-off percentage	• Yes/No Method: Experts asked if borderline student can perform the item (yes/no) • Extended Method: mix of constructed- and selected-response items; experts estimate the scale points they believe borderline examinees will obtain on each constructed-response item	• Experts categorize each item according to relevance and difficulty • Calculated score compared to a matrix to determine the probability of borderline student performing item correctly • Uses a cut-off mark for each exam based on the performance of students in relation to defined standard • Experts make judgment on individual exam items, NOT students
Orientation	Item	Item	Item

Test-centered standard setting-multiple choice exams only [2,6]

Characteristics of standard setting method	**Nedelsky**
Used for written assessments	• Panel of experts review each item and identify options that minimally-competent students should be able to eliminate as incorrect. • Minimum Passing Levels for that item is reciprocal of number of remaining options • Overall cut score determined by averaging the probability for all items
Orientation	Item

Test-centered Example: Angoff

	Item 1	Item 2	Item 3	Item 4	Item 5	Means
Rater 1	90	90	100	100	100	**96**
Rater 2	60	80	50	60	70	**64**
Rater 3	90	70	80	80	100	**84**
Rater 4	70	60	70	80	90	**74**
Rater 5	90	60	90	40	80	**72**
Mean	**80**	**72**	**78**	**72**	**88**	**78**

Passing score= 78%

Examinee-centered standard setting methods[6]

Method Characteristics	**Borderline Groups**	**Contrasting Groups**	**Bookmark Method[1,2]**
Use for clinical type assessments	• Global rating for each station by evaluator used to allocate examinees into 3 groups (passing, borderline, failing) • Cut-off score is the mean score for borderline group	• Examinees are allocated into "passing and failing groups“ by evaluator • Mean score calculated; cut-off is midpoint between means of passing/failing groups	• Measured items ordered in level of anticipated difficulty (easy to hard) • Round 1: Judges identify initial evidence threshold of competence • Round 2: Judges review ratings from round 1 and compare differences • Round 3: Evaluation of median ratings of all groups and pass/fail points • Overall median used to determine passing score using item response theory
Orientation	Mixed	Person-centered	

Borderline Group Example[5]

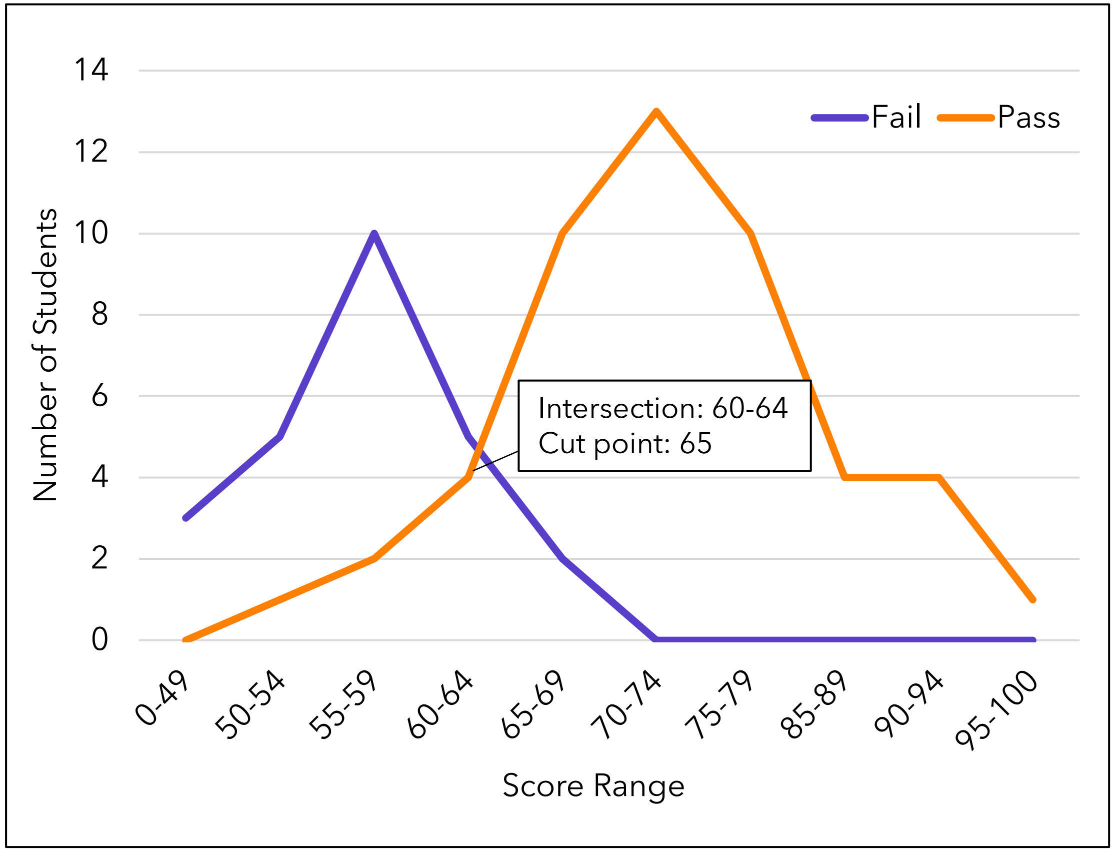
Student	OSCE Score	Rating
1	75	Clear Pass
2	83	Superior
3	75	Clear Pass
4	100	Superior
5	75	Clear Pass
6	92	Superior
7	92	Superior
8	83	Superior
9	83	Superior
10	60	Clear Fail
11	40	Clear Fail
12	50	Clear Fail
13	60	Clear Fail
14	70	Clear Pass
15	80	Superior
16	70	Clear Pass
17	90	Superior
18	64	Borderline
19	50	Clear Fail
20	57	Clear Fail
21	43	Clear Fail
22	64	Borderline
23	71	Clear Pass
24	71	Clear Pass
25	71	Clear Pass
26	89	Superior
27	79	Clear Pass
28	64	Borderline
29	64	Borderline
30	89	Superior
31	58	Clear Fail
32	74	Clear Pass
33	74	Clear Pass
34	95	Superior
35	70	Clear Pass
36	80	Superior
37	56	Clear Fail
38	75	Clear Pass
39	69	Borderline
40	50	Clear Fail
41	81	Superior
42	63	Borderline
43	50	Clear Fail
44	68	Borderline
45	68	Borderline
46	89	Superior
47	84	Superior
48	94	Superior
49	69	Borderline
50	75	Clear Pass
51	92	Superior

Borderline Group Median Score: 64

Contrasting Group Example[5]

Score Ranges and Frequencies

	Examiner Decision			
Score range	Fail	Pass	Total	Pass Rate
0-49	3	0	3	100%
50-54	5	1	6	96%
55-59	10	2	12	88%
60-64	5	4	9	72%
65-69	2	10	12	59%
70-74	0	13	13	43%
75-79	0	10	10	26%
85-89	0	4	4	12%
90-94	0	4	4	7%
95-100	0	1	1	1%



Intersection: 60-64
Cut point: 65

Audience Response Question

How do you set cut scores (e.g., determine pass/fail) for high-stakes assessment at your school (or within your program)?

a. Grade-based method (e.g., 70% to pass assessment)

b. Norm-referenced (e.g., bottom 10% of performers remediated assessment)

c. [Modified] Angoff method (e.g., use of experts to gather an overall score)

d. Borderline groups (e.g., identify score based on borderline students)

e. Mixture of above methods

f. None of the above methods

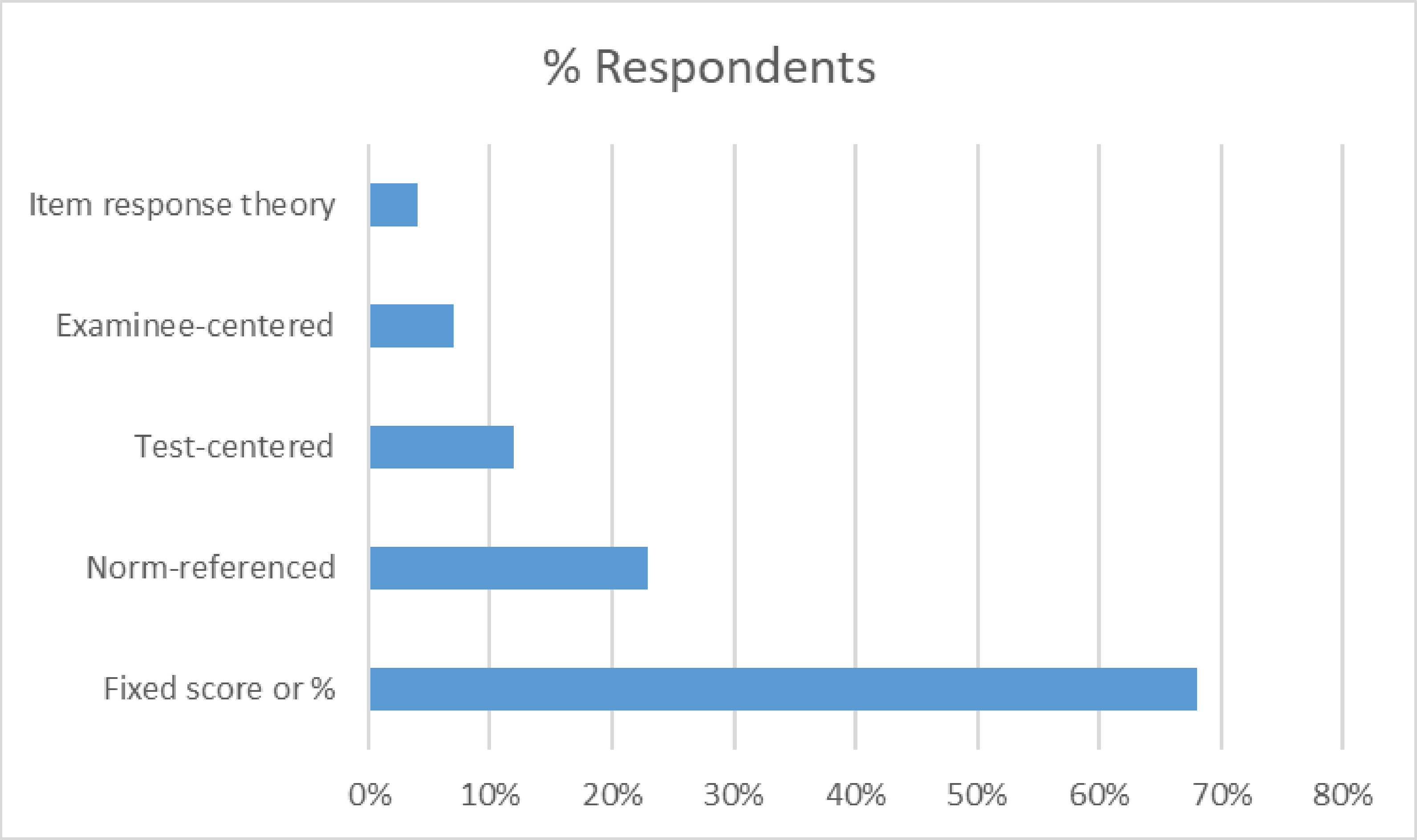
LMU

LINCOLN MEMORIAL UNIVERSITY

Google form responses: https://forms.office.com/r/p8jCP9USzz

Survey-Passing Score Determination

How was the passing score determined for the progression assessment developed by the school? (Select all that apply)



LMU
LINCOLN MEMORIAL UNIVERSITY

Considerations when selecting a standard-setting method

LMU

LINCOLN MEMORIAL UNIVERSITY

Choosing a Method: Resources

- Experts to serve as panelists
  - Test-centered methods such as Angoff require 10-15 judges[9]
  - Some examinee-centered methods require [prior] knowledge of actual examinees' ability
- Experienced facilitator(s)
- Time and expertise for data analysis
  - Bookmark method involves item difficulty analyses using IRT

Choosing a Method: Time and Timing

- Time
  - All methods require time from faculty and staff
  - Some methods are simpler and require less time, such as Borderline Group
- Timing
  - Test-centered methods can provide cut score *before* whereas examinee-centered provide cut score *after* the assessment

LMU

LINCOLN MEMORIAL UNIVERSITY

Choosing a Method: Access to Information[10]

- Test-centered methods require exam items or criteria and tasks be provided to panelists
- Some examinee-centered methods depend upon panelists having access to actual student performance information

Choosing a Method: Sample Size

- Depending on the assessment, placement in curriculum, student preparedness, and cohort size, the number of borderline students may be small
- A small *N* is mainly a challenge with examinee-centered methods that rely on actual performance data
  - Can lead to an unstable cut score and incorrect classification of students (threat to validity)

Choosing a Method: Subjectivity

- All methods involve identifying borderline students and making judgments about expected or actual performance
- Bookmark method provides greater objectivity by using difficulty estimates produced from IRT analysis and judges' ratings[10]

Discussion on standard-setting challenges

LMU

LINCOLN MEMORIAL UNIVERSITY

Active Learning Activity

- Using the provided handout, identify 1-2 standard settings methods that could be used within your program
- Determine 1 challenge that you would need to address and a possible solution to this challenge

DISCUSSION: Challenges and strategies to overcome challenges

Challenge	Strategies to overcome challenges
Use of experts and/or experienced facilitators (Resources)	
Time to conduct pre-analysis work (i.e., collect thoughts and opinions of experts)	
Allowing experts (not faculty) to access student performance data	
Determining what is a borderline/average performance	

References

1. Zieky, M. & Perie, M. (2006). A primer on setting cut scores on tests of educational achievement. *ETS*. https://www.ets.org/research/policy_research_reports/publications/publication/2006/dbkw

2. Cizek, G.J. (2006). Standard setting. In S.M. Downing, T.M. Haladyna, M.R. Raymond, S. Lane. (Eds.) *Handbook of test development* (pp. 225-257). L. Earlbaum.

3. Hubley, A.M. & Zumbo, B.D. (2011). Validity and the consequences of test interpretation and use. *Soc Indic Res* 103, 219. https://doi.org/10.1007/s11205-011-9843-4

4. Brennan, R.L. (2006). Perspectives on the evolution and future of educational measurement. In R.L. Brennan (Ed.), *Educational measurement* (4th Ed.) (pp. 1-16). Westport, CT: American Council on Education/Praeger.

5. 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. https://doi.org/10.3109/0142159X.2013.853119

6. Hays, R. (2015). Standard setting. *Clinical Teacher 12*, 226-230. https://doi.org/10.1111/tct.12395

7. Erwin, T.D., Wise, S.L (2001). Standard setting. *New directions for institutional research, 110*, 55-64.

References

8. *Item Response Theory*. (2019). Population health methods. Columbia University Mailman School of Public Health. Retrieved June 1, 2022, from https://www.publichealth.columbia.edu/research/population-health-methods/item-response-theory#:~:text=The%20item%20response%20theory%20(IRT,outcomes%2C%20responses%20or%20performance).

9. Fowell, S.L., Fewtrell, R., & McLaughlin, P.J. (2008). Estimating the minimum number of judges required for test-cetenred standard-setting on written assessments. Do discussion and iteration have an influence? *Advances in health sciences education, 13*, 11-24. DOI 10.1007/s10459-006-9027-1

10. Pitoniak, M.J. & Morgan, D.L. (2011). Setting and validating cut scores for tests. In C. Secolsky & D.B. Denison, *Handbook on measurement, assessment, and evaluation in higher education* (2nd Ed.). New York, NY: Routledge.

Questions?

LMU

LINCOLN MEMORIAL UNIVERSITY

Presenters

Mike Rudolph, PhD
Assistant Dean of Academic Affairs and Assistant Professor
Lincoln Memorial University School of Medical Sciences
michael.rudolph@lmunet.edu

Jill Augustine, PharmD, PhD, MPH
Director of Assessment and Assistant Professor
Mercer University College of Pharmacy
Augustine_jm@mercer.edu

Justine Gortney, PharmD, BCPS
Director of Assessment, Division of Pharmacy and Associate Professor
Eugene Applebaum College of Pharmacy and Health Sciences
Wayne State University
justine.gortney@wayne.edu

LMU
School of Medical Sciences
LINCOLN MEMORIAL UNIVERSITY

Correct or Incorrect? Objective Structured Clinical Exam (OSCE) Standard Setting and "Grading"

Session Worksheet
IUPUI Assessment Institute

Rudolph M, Augustine J, & Gortney JS

1) Which of the standard-setting options listed below currently aligns with (or would best align with) an OSCE or other performance-based assessment for your program?

Standard-setting Options	
Test-centered Options	**Examinee-centered Options**
Angoff	Borderline Group
Modified-angoff	Contrasting Group
Ebel	Bookmark

2) Based on the standard-setting option selected above, describe at least one challenge utilizing this particular standard-setting option and provide one or more potential solutions to address this challenge.

Challenge(s) with chosen standard - setting option	Potential solutions for overcoming standard setting challenges