Standard Setting

in Medical Education

A Comprehensive Overview for Fellows of the
West African College of Physicians

Wole Adebiyi
Department of Medicine
College of Medicine
University of Ibadan
University College Hospital
Ibadan

Learning Objectives

By the end of this session, participants will be able to:

1. Define standard setting and explain its central importance in postgraduate specialist certification
2. Distinguish between norm-referenced and criterion-referenced approaches and justify the use of criterion-referencing in WACP examinations
3. Describe and compare the principal standard-setting methods: Angoff, Ebel, Hofstee, Borderline Group, and Borderline Regression
4. Select the appropriate method for each assessment format used in WACP examinations
5. Outline the process for convening and training a standard-setting panel, including definition of the minimally competent candidate
6. Apply principles of validity, reliability, and defensibility to evaluate and improve standard-setting practice within the faculties

SECTION 1

What is Standard Setting?

Definitions, concepts, and the stakes in specialist certification

Definition and Core Question

Standard setting is the process of determining the minimum level of performance required to demonstrate competence in an assessment.

The fundamental question:

How good is good enough: and how do we know?

For the WACP:

Certifying a specialist authorises independent practice, supervising residents, and high-stakes clinical decisions across 14 member nations.

Two errors matter:

A false pass threatens patient safety. A false fail is an injustice to a competent candidate and reduces the specialist workforce West Africa urgently needs.

Norm-Referenced vs Criterion-Referenced Standards

Norm-referenced (relative)

- Pass mark set relative to cohort performance
- Bottom 10%, or mean minus 1 SD: common examples
- Standard moves with each cohort of candidates
- A strong cohort = higher bar; a weak cohort = lower bar
- No stable relationship to clinical competence
- Some competent candidates will always fail
- **NOT appropriate for licensing/certification**

Criterion-referenced (absolute)

- Pass mark anchored to required competency level
- Independent of how the rest of the cohort performs
- Standard is stable across examination cohorts
- Hypothetically: all candidates could pass or all could fail
- Stable, defensible relationship to clinical competence
- Methods include: Angoff, Borderline Group, BRM
- **The appropriate approach for WACP certification**

The arbitrary 50% pass mark has no defensible relationship to clinical competence: it must be replaced.

SECTION 2

Standard Setting Methods

Test-centred and examinee-centred approaches

Overview: Two Families of Methods

All criterion-referenced: differing in what the judges evaluate

Test-centred methods

Judges evaluate items

- Modified Angoff
- Original Angoff (probability)
- Ebel method
- Hofstee compromise

Examinee-centred methods

Judges evaluate performances

- Borderline Group Method (BGM)
- Borderline Regression Method (BRM)
- Band descriptor approach
- Global rating + checklist hybrid

The Angoff Method

The most widely used standard-setting procedure for written examinations

The procedure

- Expert panel (8–12 judges) reviews every examination item
- For each item, judges ask: *"What proportion of minimally competent candidates would answer this correctly?"*
- Modified (Yes/No) variant: binary decision per item: reduces cognitive burden
- Each judge's item estimates are summed to give a candidate cut score
- Final pass mark = mean/median of all judges' estimates across the panel
- Can be applied prospectively: before candidates sit the exam
- Transparent and auditable: every judgement is recorded

Strengths

- Prospective: pre-exam
- Transparent, documented
- Usable by non-psychometricians

Limitations

- Logistically demanding
- Judges must be trained carefully
- Borderline candidate definition is challenging
- Tends to produce lower cut scores than examinee-centred methods

Best suited for: MCQ / EMQ components of written examinations

The Ebel Method

Adds a relevance dimension to item-level standard setting

Items are classified on two dimensions simultaneously:

	Essential	Important	Acceptable	Questionable
Easy	% correct estimate	% correct estimate	% correct estimate	% correct estimate
Medium	% correct estimate	% correct estimate	% correct estimate	% correct estimate
Difficult	% correct estimate	% correct estimate	% correct estimate	% correct estimate

Shaded cells = high relevance + lower difficulty (most critical items for setting the standard)

The Ebel method forces judges to consider whether items are worth including at all: items rated 'questionable' in relevance arguably should not be in a certification examination.

The Hofstee Compromise Method

Bridges absolute standard setting and awareness of actual candidate performance

Judges answer four questions: defining a rectangle of acceptable outcomes:

Q1	What is the minimum acceptable pass rate?
Q2	What is the maximum acceptable pass rate?
Q3	What is the minimum acceptable pass mark?
Q4	What is the maximum acceptable pass mark?

How the cut score is determined:

The four values define a rectangle on a graph. The actual pass mark is the point where the real score distribution curve intersects the diagonal of this rectangle.

Most useful when transitioning from norm-referenced to criterion-referenced standard setting.

The Borderline Group Method (BGM)

Examinee-centred standard setting for OSCEs and clinical assessments

The procedure

1. Each OSCE examiner rates every candidate on a global rating scale: Fail / Borderline Fail / Borderline Pass / Pass / Clear Pass
2. Independent of the global rating, the examiner also scores the candidate on the structured checklist
3. Pass mark for each station = mean checklist score of all candidates rated 'Borderline' (combining Borderline Fail and Borderline Pass)
4. Overall OSCE pass mark = average of station-level pass marks
5. The borderline candidate is real: identified by experienced clinicians in the actual examination, not hypothetical

Strengths

- Strong face validity with clinical examiners
- Intuitive: no separate panel meeting
- Anchored in real observed performance

Limitations

- Unstable when few borderline candidates at a station
- Risky with small WACP cohorts
- Cannot be determined prospectively
- Superseded by BRM in many contexts

Best suited for: manned OSCE stations: direct observation of clinical performance

The Borderline Regression Method (BRM)

The most statistically robust examinee-centred approach

How it works

- Examiner completes both a structured checklist AND a global rating for each candidate at each station
- Checklist scores from ALL candidates are regressed on global rating scores: producing a linear equation
- Pass mark = checklist score predicted by the regression at the borderline value of the global rating scale
- Uses all candidates' data: not just those rated borderline: greatly improving statistical stability
- NPMCN-approved for both OSCE stations and essay questions

BRM vs BGM

	BRM	BGM
Data used		
Statistical stability	**High**	Lower
Small cohorts	**Suitable**	Risky
Prospective?	**No**	No
Preferred for essays?	**Yes**	No
NPMCN approved?	**Yes**	Yes

Best suited for: OSCE stations and essay questions: preferred over BGM when cohort size is small

Matching Methods to Assessment Formats

The right method for the right examination component

Assessment format	Recommended method	Approach	Notes
MCQ / EMQ	Modified Angoff (Yes/No)	Test-centred	*Prospective; panel meets before exam*
Essay / MEQ	Borderline Regression (BRM)	Examinee-centred	*Uses checklist + global rating*
OSCE: observed stations	BGM or BRM	Examinee-centred	*BRM preferred for small cohorts*
OSCE: unobserved stations	Modified Angoff	Test-centred	*No real-time examiner observation*
Dissertation / viva	Band descriptors + global rating	Examinee-centred	*Holistic, qualitative approach*
Transitional / legacy systems	Hofstee compromise	Compromise	*Bridges old and new approaches*

Source: Adapted from NPMCN standard-setting framework (2024) and international best practice guidelines

SECTION 3

The Standard-Setting Process

Panel composition, training, and the step-by-step procedure

The Standard-Setting Panel: Composition and Training

Panel size

8–12 judges is recommended; as few as 5 and as many as 20 have been used, but 10 is the most widely supported number for reliable outcomes.

Panel composition

Fellows from multiple member nations; range of seniority; diverse training backgrounds; at least one recently certified Fellow who can speak to current training experience.

Structured training

Training must include: criterion vs norm-referenced thinking; definition of the borderline candidate; practice rounds with sample items; discussion of outlier estimates.

The reality check

After Round 1, judges see the actual score distribution. If the derived cut score would result in 0% or 95% passing, re-calibration is appropriate: not a compromise of rigor.

Conceptual drift

Judges' concept of minimal competence shifts over long sessions due to fatigue, item exposure, and discussion effects. Session management: breaks and anchoring reminders: is essential.

The West African context

The borderline candidate must be defined with explicit reference to WACP practice settings: competencies required across tertiary and district hospital contexts in 14 member nations.

The Standard-Setting Process: Step by Step

1	2	3	4	5	6
Select and brief the panel	Define the borderline candidate	Train judges in the method	Round 1 independent ratings	Discussion and reality check	Calculate and ratify cut score
8–12 judges; multi-country WACP representation; define selection criteria; provide pre-reading on the examination blueprint	Formal, written, behavioural description of minimal competency	Practice Angoff rounds with retired items; calibration exercises; discussion of conceptual drift and norm vs criterion thinking	Each judge rates every item independently, without discussion; prevents anchoring bias and group conformity effects	Review and discuss outlier judgements; show actual score distribution; judges may revise estimates in light of evidence	Mean/median of panel estimates; Faculty Board ratification; full documentation archived for accreditation and appeals

Findings feed back to refine subsequent cycles: standard setting improves with institutional experience

SECTION 4

Validity, Reliability & Defensibility

The three pillars of sound standard-setting practice

The Three Pillars of Sound Standard Setting

Validity

Does the standard measure what it claims: the minimum competence required for safe specialist practice in West Africa?

Key threats: construct under-representation (testing only pharmacology when holistic management is the construct); using pass marks calibrated only at tertiary centres.

The standard must reflect competencies required across all WACP practice settings: not just major teaching hospitals.

Reliability

Is the cut score consistent and reproducible? Reliability is threatened by: judge variability, small numbers of borderline candidates (BGM), and insufficient training.

Measure: intraclass correlation among Angoff judges (target ≥0.70). For BRM: root mean square error of the regression.

The BRM is generally more reliable than BGM because it uses all candidates' data, not just borderline cases.

Defensibility

Can the cut score be justified to candidates, professional bodies, accreditation authorities: and courts if necessary?

Required documentation: rationale for method choice; panel composition and training records; individual and aggregate judge estimates; reality check notes; final ratification decision.

For WACP: Faculty Board ratifies the cut score; College Council's Education Committee provides oversight.

Psychometric Considerations After Setting the Cut Score

Item analysis

Post-examination review of each question: difficulty index (% answering correctly), discrimination index (correlation with total score), point-biserial correlation. Items with very low discrimination should be reviewed or retired.

Examination reliability

Cronbach's alpha measures internal consistency of the full paper. High alpha is necessary but not sufficient: many candidates scoring near the cut score means many erroneous decisions even with high alpha.

Standard error of measurement

The SEM quantifies uncertainty around any individual score. High-stakes pass-fail decisions should always be interpreted in light of SEM: scores within 1 SEM of the cut score require careful governance.

Post-hoc pass rate review

Is the resulting pass rate consistent with historical trends? Unexpected spikes in failure warrant investigation: check for harder items, weaker cohort, or standard-setting error before ratifying results.

Longitudinal comparability

For multiple examination diets per year, statistical equating techniques ensure the standard is genuinely comparable across different examination versions and cohorts.

Feedback to question writers

Item analysis data should flow back to the examination question bank: poor-performing items are flagged for revision, and patterns of low difficulty identify curriculum gaps requiring attention.

SECTION 5

Standard Setting in the WACP Context

Challenges, opportunities,

WACP-Specific Challenges and Opportunities

Multi-faculty complexity

Six faculties across 14 nations require faculty-specific protocols within a College-wide methodology and governance framework. Each Faculty Board should develop its own standard-setting procedure.

Examination level differentiation

Primary (MCQs) → Angoff. Membership and Fellowship (OSCEs, essays, clinical) → BRM/BGM. Method-specific protocols are needed for each level of the WACP examination system.

Multi-country training context

The borderline candidate must reflect competencies required across the full range of WACP practice settings: from University College Hospital Ibadan to a district hospital in Sierra Leone.

Building examiner capacity

The Doctors as Educators and ToT programmes are the natural vehicles for standard-setting training. Integrating practical Angoff workshops into existing faculty development infrastructure.

Resource-conscious implementation

BRM (essays/OSCEs) and modified Angoff (MCQs) offer the best balance of rigor and operational feasibility. Neither requires external psychometric consultants: trained faculty can implement both.

Pilot and iterative approach

Introduce criterion-referenced standard setting via one faculty pilot, gather evidence across multiple diets, then lead a College-wide rollout: following the NPMCN model of incremental implementation.

Standard Setting in Competency-Based Medical Education

Looking ahead: entrustment-based standards for WACP

The shift toward CBME: already begun in the WACP Faculty of Family Medicine: changes the standard-setting question from 'Did this candidate pass?' to 'Is this candidate ready for unsupervised practice?'

Traditional examinations	CBME / EPA-based assessment
Discrete pass/fail decision on a test	Entrustment decision across multiple EPA observations
Single high-stakes examination event	Longitudinal workplace-based assessment
Angoff / BGM / BRM methods	Entrustment scale (1–5) with threshold standards
Numeric cut score	Minimum number of observations at specified level

The EPA entrustment scale requires its own standard-setting logic: what level, across how many observations, constitutes readiness for unsupervised practice? This is the WACP's next frontier.

Summary Learning Points

Key concepts to carry forward from this session

1

Standard setting is a moral responsibility

For WACP, the pass mark determines who is certified to practise independently: two errors are possible, and both carry serious consequences for patients and candidates.

2

Criterion-referenced is the only defensible approach

Norm-referenced methods (including arbitrary 50%) have no stable relationship to clinical competence and cannot be defended before accreditation bodies or courts.

3

Method must match format

MCQs → modified Angoff. OSCEs and essays → BRM (preferred) or BGM. Mismatching methods to formats produces unreliable cut scores.

4

The borderline candidate is the conceptual key

All standard-setting methods rest on a shared understanding of minimal competency. Invest time in defining the WACP borderline candidate concretely and contextually.

5

Documentation is non-negotiable

Every step: panel selection, training, item ratings, reality check, final ratification: must be recorded. Undocumented standard setting is not defensible.

Recommendations for the WACP

A roadmap toward criterion-referenced standard setting across the College

R1

Adopt criterion-referencing as College policy

Formally retire the arbitrary 50% pass mark across all faculties and examination levels. Mandate criterion-referenced standard setting through an approved College policy document.

R2

Establish a standard-setting protocol per faculty

Each Faculty Board develops format-specific protocols: modified Angoff for MCQs; BRM for OSCEs and essays. Protocols to be approved by the Education and Research Committee.

R3

Integrate training into faculty development

Add standard-setting workshops: including live Angoff practice sessions: to the Doctors as Educators programme and all Training of Trainers retreats.

R4

Define the WACP borderline candidate formally

Develop a written, behavioural description of the minimally competent WACP Fellow for each faculty, explicitly referencing the range of West African healthcare settings.

R5

Pilot in one faculty before College-wide rollout

Select one faculty to pilot criterion-referenced standard setting over two to three examination diets. Publish outcomes internally. Use evidence to refine the College-wide protocol.

R6

Build a post-examination analysis infrastructure

Establish routine item analysis, pass rate review, and reliability reporting for every examination diet. Integrate findings into question bank governance and faculty review cycles.

R7

Prepare for CBME integration

Commission the Education and Research Committee to develop entrustment-based standard-setting guidance for faculties implementing CBME curricula and EPA-based

R8

Document, govern, and review continuously

Maintain an archive of standard-setting records for every examination cycle. Conduct biennial reviews of methods, outcomes, and alignment with current international best

The greatest contribution of standard setting to the WACP is not a number: it is a culture of evidence-based professional accountability.

Thank you for listening