A COMPANION ESSAY · NAIROBI · 2026

The Angoff Panel for Testing Clinicians

Where the pass mark comes from, and why an arbitrary 50% cannot be defended when a certificate is a safety claim.

Dr Neal Aggarwal

Medicine · Surgery · Engineering · Information Technology · Artificial Intelligence

12 AUGUST 2026

One line on one slide of the Level 1 deck reads: *Standard set by modified Angoff panel (No arbitrary 50% pass rate).* It is the least glamorous sentence in the whole curriculum and possibly the most consequential. **This essay unpacks it completely** — what a cut score actually is, why 50% is indefensible and norm-referencing is worse, who the borderline candidate is and how you build one, the mechanism step by step with a full worked panel whose arithmetic you can check, what "modified" really means, what the 2025 meta-analysis of 91 studies says about which variant to choose, and precisely where Angoff stops working.

Contents

The final slide of the Level 1 facilitator deck is called **Assessment Architecture**. In the Knowledge Assessment column, between “40-item invigilated MCQ” and the note about Discernment carrying 40% of the weight, sits this:

> **Standard set by modified Angoff panel (No arbitrary 50% pass rate).**

It is the least glamorous line in the entire curriculum. It is also, I think, one of the two or three most consequential, because it is the line that decides who is certified and who is not — and more importantly, it decides *on what basis* that decision can be defended when someone challenges it.

I want to unpack it completely. Not the textbook summary, which most readers of this blog will already have, but the whole thing: what a cut score actually is, why the obvious answers are wrong, who the borderline candidate is and how you build one, the mechanism in operational detail, a fully worked panel with the arithmetic exposed, what “modified” genuinely means as against what everybody says it means, what the recent evidence says about which variant to choose, where the method fails, and exactly where it stops applying.

---

1. What a cut score actually is

Start with the thing that is easy to miss because it sounds like a technicality.

A cut score is not a measurement. It is a policy decision about acceptable risk, expressed in whatever units the test happens to produce.

Nothing in the candidate’s performance tells you where the line goes. The test gives you a number between 0 and 40. Where you draw the line across that scale is a judgement about how much incompetence you are prepared to certify — and, symmetrically, how many competent people you are prepared to fail. Every cut score trades those two errors against each other. There is no value-free place to stand.

This has a liberating consequence and a demanding one. The liberating one: you are allowed to make the judgement, because someone must. The demanding one: because it is a judgement rather than a discovery, the *only* thing that makes it defensible is the quality and transparency of the process that produced it.

That is the whole game. Everything below is machinery for making a judgement defensible.

The two answers that do not work

“Fifty per cent.” Why fifty? Because it is half. That is not a reason, it is a coincidence of the decimal system. Worse than arbitrary, it is *unstable*: the pass mark is fixed but the difficulty of the paper is not. Write a harder paper this year and you have silently raised

the standard without deciding to. Write an easier one and you have lowered it. The number stays at 50 and the thing it means drifts underneath you. A fixed percentage is a standard that changes every time you change the exam, which is the precise opposite of what a standard is for.

This is why the slide says what it says. “No arbitrary 50% pass rate” is not a swipe at a straw man — it is the default that most training programmes actually use, including many that would describe themselves as rigorous.

“The top 70% pass.” This is **norm-referencing**: the standard is defined relative to the cohort’s performance rather than to the content. It is administratively wonderful. It guarantees a predictable supply of certified practitioners, which is exactly what a ministry planning workforce numbers wants to hear.

It is also indefensible for a safety qualification, and the reason is stark. Norm-referencing guarantees that a fixed proportion passes **regardless of whether anybody is competent**. In a weak year you certify people who should not be certified. In a strong year you fail people who are safe. The standard floats, and what it floats on is the accident of who happened to sit the paper that month. A patient’s entitlement to a competent clinician does not vary with the strength of that clinician’s cohort.

Criterion-referencing is the alternative: the standard is derived from the content, item by item, from a judgement about what a minimally competent practitioner must be able to do. It can be set before a single candidate sits the paper. Angoff is the dominant criterion-referenced method for knowledge tests, and has been for fifty years.

---

2. The borderline candidate

Everything in an Angoff panel hangs on one construct, and if the construct is vague the whole exercise is theatre.

The **borderline candidate** — also called the minimally competent, minimally acceptable, or just-qualified candidate — is the person sitting exactly at the boundary of acceptable practice. Not a good candidate. Not a weak one. Someone who could plausibly go either way, and about whom a reasonable examiner would say: *this is the least I am willing to certify.*

Judges are not being asked what they think of the items. They are being asked to predict the behaviour of a specific hypothetical person. If five judges hold five different people in mind, they are answering five different questions and averaging the answers produces a number that means nothing.

So the first hour of a properly run panel is not spent on the paper at all. It is spent constructing that person together, out loud, in writing. The output is a set of **anchor descriptions** — concrete statements of what the borderline candidate can and cannot do.

For Level 1 Clinical AI Foundations, mine would read something like this:

> **The borderline candidate.** They can name the three modalities and will usually classify a clear-cut case correctly, but hesitate and sometimes err when a case sits between augmentation and agency. They know identifiable patient data must not leave a controlled system and will say so, though they may not immediately recognise a disguised instance of it. They can recite the independent-impression rule and understand why it exists; under time pressure they may describe writing the impression *after* consulting the system, and not notice that this defeats the point. They will catch a frank dosing error in AI output. They will often miss an error of omission. They know a differential generated with AI assistance must be documented as such, but their phrasing tends to be vague about who made the final decision. They know the Data Protection Act 2019 and the Digital Health Act 2023 exist and roughly what they require; they cannot cite sections. They are safe under supervision. They are not yet safe unsupervised — which is precisely what a Level 1 certificate claims, and no more.

Write that, agree it, and give every judge a copy in front of them for the whole exercise. Anchor statements are not a formality; there is published concern that poorly written ones import their own distortions, so they are worth drafting as carefully as the items themselves.

A note on the word "minimally". It does real work and it is frequently softened in practice. Panels drift toward describing a *good* candidate because that is who they enjoy imagining, and every step in that direction pushes the cut score up. Guarding against that drift is the chair's main job.

3. The mechanism, step by step

Who sits on the panel

Annex C specifies **not fewer than eight practising clinicians spanning the cadres in the cohort**. Both halves of that matter.

Not fewer than eight is about stability: with three judges, one outlier moves the standard substantially. There is now direct evidence on panel size, which I will come to in section 7, and it is not quite what you would expect.

Spanning the cadres is about whose picture of "minimally competent" gets encoded. A panel of consultants estimating what a borderline **nurse** must know will produce a standard reflecting consultants' beliefs about nurses. Those beliefs are frequently wrong in both directions — over-estimating familiarity with the law, under-estimating clinical pattern recognition. Because the Level 1 core is taught mixed-cadre and assessed on

one common paper, the panel must be mixed too, or the standard silently becomes a medical standard applied to everyone.

What each judge does

For every item on the paper, independently and without discussion, each judge answers exactly one question:

> **Of 100 borderline candidates, how many would answer this item correctly?**

The answer is recorded as a proportion — 0.65, say. Not "is this a good item", not "how hard is this", not "would I get it right". A prediction about a specific population's performance on a specific item.

Two constraints worth stating explicitly to a panel:

- The floor is not zero, it is **chance**. On a five-option single-best-answer item, a candidate who knows nothing scores 0.20 by guessing. An estimate of 0.05 is incoherent. Panels violate this constantly.
- The ceiling is not one. If every borderline candidate would get an item right, the item is discriminating nothing and belongs in the bin, not the paper.

How the numbers combine

Each judge's estimates are averaged across items to give **that judge's cut score**. Those are averaged across judges to give **the panel's cut score**. That number, expressed as a percentage of the total marks, is the pass mark.

The arithmetic is deliberately trivial. The defensibility lives entirely in the process, the documentation, and the construct — not in the maths.

---

4. A worked panel

Abstractions about standard setting are easy to nod along to and hard to actually hold. So here is a complete panel: eight items, five judges, two rounds, with every estimate exposed so you can check the arithmetic yourself.

The items are drawn from across the Level 1 units, with the difficulty spread you would want on a real paper — some near-universal, some genuinely discriminating.

Round 1 — independent estimates

#	Unit	What the item asks	A	B	C	D	E	mean
1	1.3	Identify which listed task is non-delegable	0.88	0.80	0.92	0.75	0.85	**0.84**
2	1.1	What fluent, correctly formatted output does and does not establish	0.75	0.68	0.82	0.60	0.72	**0.71**
3	1.2	Classify a pre-sorting triage assistant: automation, augmentation or agency	0.55	0.45	0.68	0.40	0.57	**0.53**
4	1.6	Which action breaches the Data Protection Act 2019	0.90	0.82	0.93	0.80	0.88	**0.87**
5	1.5	At which step of the independent-impression rule must the clinician write	0.85	0.78	0.88	0.71	0.82	**0.81**
6	1.5	Correct response to a differential that omits TB in a wasting patient	0.62	0.55	0.73	0.45	0.60	**0.59**
7	1.6	Which documentation phrasing correctly discloses AI involvement	0.72	0.62	0.78	0.58	0.68	**0.68**
8	1.5	Classify a behaviour: under-trust, calibrated trust, or automation bias	0.73	0.60	0.76	0.62	0.69	**0.68**
		Judge cut score	**75.0**	**66.2**	**81.3**	**61.4**	**72.6**	**71.3%**

Panel cut score: **71.3%**. Between-judge standard deviation: **7.7 points**.

Look at that spread before anything else. Judge C would set the bar at 81.3%; Judge D at 61.4%. **Twenty percentage points apart on the same eight items.** If you stopped here and published 71.3% as your standard, you would be publishing the average of two incompatible beliefs about what competence means. The mean of a disagreement is not a consensus.

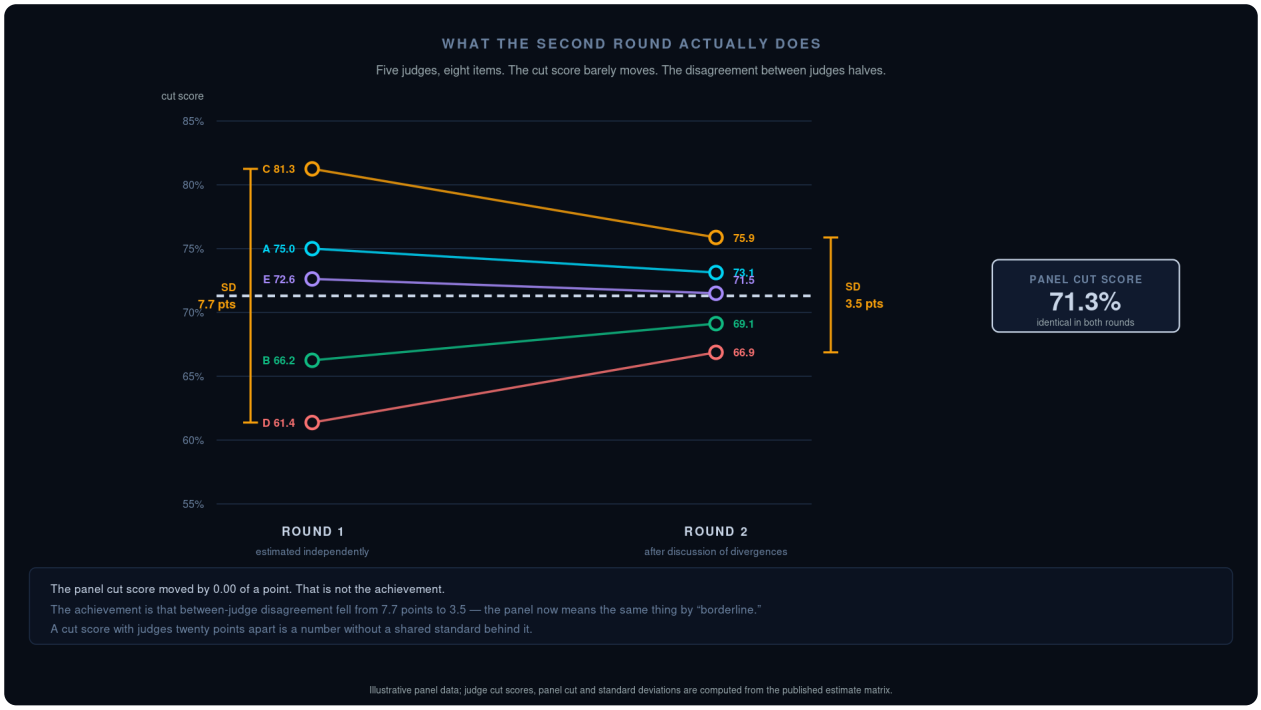
That is what round two is for.

Round 2 — after discussion of divergences

Judges see the spread of estimates item by item — usually anonymised — and the items where they most disagree are discussed. Then everyone re-estimates independently. Nobody is asked to concede.

#	Unit	A	B	C	D	E	mean	change
1	1.3	0.88	0.82	0.92	0.80	0.85	**0.85**	+0.01
2	1.1	0.74	0.70	0.79	0.66	0.72	**0.72**	+0.01
3	1.2	0.50	0.47	0.52	0.45	0.51	**0.49**	**−0.04**
4	1.6	0.90	0.85	0.92	0.84	0.88	**0.88**	+0.01
5	1.5	0.85	0.80	0.85	0.78	0.82	**0.82**	+0.01
6	1.5	0.58	0.56	0.64	0.53	0.58	**0.58**	−0.01
7	1.6	0.72	0.66	0.72	0.63	0.68	**0.68**	+0.01
8	1.5	0.68	0.67	0.71	0.66	0.68	**0.68**	0.00
	Judge cut score	**73.1**	**69.1**	**75.9**	**66.9**	**71.5**	**71.3%**	

Panel cut score: **71.3%**. Between-judge standard deviation: **3.5 points**.



The cut score is identical in both rounds. The disagreement between judges falls by 55%. That is the point of the second round.

What just happened

The panel cut score did not move. Not approximately — exactly. 71.3% both times, to the last decimal the arithmetic will give.

If the purpose of round two were to find a better number, round two accomplished nothing at all and you should stop running it.

The purpose of round two is not the number. It is to establish that **the panel means the same thing by "borderline."** A cut score of 71.3% derived from judges who are twenty points apart is a number with no shared standard behind it. The same 71.3% derived from judges within seven points of each other is a number that five experts, having argued, would each independently defend. Those are entirely different objects that happen to be printed identically.

The between-judge SD is therefore the statistic to report and publish alongside the cut score, and I would go further: **a panel whose spread does not narrow between rounds has failed and should not be used.** Either the construct was never shared, or the discussion was pro forma.

The one item that genuinely moved

Item 3 — classifying a pre-sorting triage assistant as automation, augmentation or agency — fell from **0.53 to 0.49**, the only item to move by more than 0.01.

That is the panel discovering something. In discussion, judges realised the item is harder than it looks: the distinction between augmentation and agency turns on whether the human is in the loop *for this specific patient*, and a triage assistant that pre-sorts a queue is agency wearing augmentation's clothes. That is exactly the confusion the Unit 1.2 slide flags as the catastrophic risk — a clinician operating in agency mode while believing they are in augmentation mode.

So a borderline candidate would find it harder than the panel first assumed. The estimate came down. This is the mechanism working as designed: not consensus for its own sake, but a shared re-reading of what an item actually demands.

---

5. What "modified" really means

Here is where I have to correct something — including a version of it I have said myself.

The usual account goes: *Angoff's 1971 original asked judges a yes/no question, and the modified version replaced it with probability estimates.* That is the story in a great many textbooks and slide decks. It is not quite right, and the real history is more interesting.

Angoff's method appears in his chapter in the 1971 edition of *Educational Measurement*. In the main text he describes a procedure he attributes to **Ledyard Tucker**: for each item, judge whether the minimally acceptable person would answer it correctly — a binary, 1 or 0.

Then Angoff adds a **footnote**. In it he suggests that instead of a binary judgement, the experts could estimate the *probability* that the minimally competent candidate answers correctly, and sum those probabilities.

That footnote became the Angoff method. The most widely used standard-setting procedure in the history of educational and licensure testing is a footnote to someone else's idea.

Which means:

- The **probability estimate is not the modification.** It is Angoff's own contribution, and arguably the more authentically "Angoff" of the two.
- The binary version survives as a distinct method, usually called **Angoff Yes/No**, and it is Tucker's.
- **"Modified" today generally refers to the procedural additions** — most importantly the iterative rounds with discussion between them, and often the provision of real candidate performance data.

I have been loose about this before and it is worth being precise, because the choice of variant turns out to matter more than the label suggests. Which brings us to the evidence.

---

6. Reality checks, and an honest tension

A common addition — the **reality check**, sometimes called empirical or normative Angoff — shows judges the actual proportion of candidates who got each item right (the item's **p-value**) *between* rounds. Judge C predicted 0.68 for item 3; the real figure last year was 0.41. Judge C now knows something.

This provokes an obvious objection, and it deserves a straight answer rather than a brush-off.

Doesn't showing judges cohort performance make the method norm-referenced by the back door?

Partly, yes. Something real is being given up. A pure criterion-referenced standard is set without reference to how anyone performed, and once you show judges the p-values you have let cohort performance into the room.

Two things make the trade worth taking, in my view.

First, the *question* judges are answering does not change. They are still asked what a **borderline** candidate would do, not what the cohort did. The p-value calibrates their estimate of item difficulty — a matter of fact they are demonstrably bad at, as section 7 shows — without changing the standard they are applying. Difficulty is empirical. Where to draw the line remains a judgement.

Second, the alternative is not purity but error. A judge who believes an item is easy when it is hard is not preserving criterion-referencing; they are making a factual mistake that propagates directly into the standard.

The safeguard is procedural: show the data **after** round one, never before, so the initial estimates are uncontaminated; make it advisory rather than a target; and record how much each judge moved. A judge who simply rewrites their estimates as the observed p-values has stopped doing Angoff and started doing norm-referencing with extra steps.

7. What the evidence actually says

Until recently you had to argue about variants from first principles. In 2025, a systematic review and meta-analysis in *BMC Medical Education* pooled **91 studies** of Angoff methods in health professions education. Its findings are directly relevant to what my slide claims, and one of them should change what Annex B says.

The variant is a primary determinant of the outcome. The authors' framing is that Angoff "is not a single technique but a flexible family of approaches with distinct outcomes." Choosing a variant is choosing a level of stringency, whether or not you realise you are choosing.

Reliability differs sharply between variants. Modified Angoff *with a reality check* achieved inter-rater reliability of **r = 0.917**. The Yes/No variant — Tucker's binary — was **r = 0.536**, described as the weakest and most variable, and confirmed by bootstrap resampling. That is not a marginal difference. It is the difference between a standard that would replicate with another panel and one that might not.

Angoff is more lenient than fixed cut scores, on average. Pooled comparisons found significantly higher pass rates under Angoff than under a conventional fixed method (**OR 7.48**). Worth sitting with if you assumed rigour means severity: replacing an arbitrary 50% with a properly derived standard will often let *more* people through, because the arbitrary number was set too high for the paper.

Mastery variants are much more stringent. Mastery Angoff produced pooled cut scores of **88.2%** (86.9% with reality check), and correspondingly low pass rates — **61.4%** for Mastery Angoff. These are appropriate where near-perfect performance is the only acceptable standard, and inappropriate almost everywhere else.

And the finding I did not expect: meta-regression showed each additional judge was associated with a **0.19-percentage-point increase in the cut score** ($p = 0.003$). Bigger panels set *higher* standards. Not dramatically — going from eight judges to twelve moves the bar by under a point — but systematically, and in a direction nobody chose. Panel size is not a neutral parameter for stability alone; it has a small directional effect on the standard itself, and it should be fixed by policy and disclosed, not settled by who was available that week.

Each additional item was associated with a 0.05-percentage-point increase in the pass rate ($p = 0.001$), which is the sampling effect you would expect: longer papers are more forgiving of a single unlucky item.

What this means for Annex B

Annex B currently specifies "modified Angoff for knowledge assessment." Given this evidence, **that is under-specified**, and I intend to amend it to read *modified Angoff with reality check*, with the p-value data released only after round one and each judge's movement recorded.

I would rather write that here, in public, than quietly change the document. The whole argument of this project is that a standard is only as good as the account you can give of it.

---

8. The limits, which are real

Angoff's dominance should not be mistaken for the method being sound in every respect. Its central assumption is empirically shaky, and the honest position is to use it while knowing that.

Judges are poor at estimating item difficulty in absolute terms. This is the big one. Impara and Plake tested classroom teachers estimating item performance for their *own students* on a test they knew — about as favourable a condition as you could construct, far better than a typical standard-setting panel. Judges could **rank-order** items by difficulty reasonably well; they could not produce accurate absolute estimates. Since Angoff sums absolute estimates, this strikes directly at the mechanism.

Panels overestimate what borderline candidates know. The systematic bias runs toward imagining a stronger candidate than the definition specifies, which inflates the cut score. Anchor statements and a vigilant chair are the mitigation; neither is a cure.

Holding one person in mind across forty items is genuinely hard. By item 30 the construct has drifted. Practical countermeasures: re-read the anchor description aloud at intervals, break the paper into blocks, and consider re-estimating a few early items at the end to measure the drift rather than pretend it did not happen.

The standard depends on who is in the room. Different panels produce different numbers from the same paper. This is not a defect to be engineered away — it is what it means for a standard to be a judgement. It is an argument for documenting composition, publishing it, and keeping it stable across diets.

And what happens when the number is inconvenient? Suppose the panel returns 82% and two-thirds of the cohort fails. The temptation to relitigate the standard is enormous and must be resisted, but the situation still needs a policy decided *in advance*:

- Whether to adjust for measurement error — many bodies subtract one **standard error of measurement** from the cut score, deliberately favouring the candidate on the grounds that the test is imprecise and the cost of failing a competent person is real.
- Whether to bound the outcome with a compromise method such as **Hofstee**, in which the panel additionally states the highest and lowest cut score it would accept and the highest and lowest fail rate it would accept, and the final standard is constrained to that region. AMEE Guide No. 18 sets this out clearly.
- What the appeals route is, and what evidence a candidate is entitled to see.

Decide those before you know the result. Afterwards is too late to be credible.

9. Where Angoff stops

This is the boundary I most want educators to take away, because it is where I see the method misapplied.

Angoff only works where items have right answers.

The question — *what proportion of borderline candidates would answer this correctly?* — presupposes a binary outcome to estimate the probability of. A forty-item single-best-answer paper supplies that. A ten-minute simulated consultation does not. You cannot sensibly ask what proportion of borderline candidates would "get the station right", because a performance is not right or wrong. It is a continuous, multi-dimensional thing scored with partial credit, in which a candidate can be excellent at three things and poor at a fourth.

People do force it, and the resulting standards do not survive scrutiny.

So the same slide specifies two instruments and, necessarily, two methods:

	Knowledge Assessment	Simulated Encounter
Instrument	40-item invigilated MCQ	10-minute standardised-patient sandbox encounter
Standard set by	Modified Angoff panel	Borderline regression
Judgement type	Item-centred, prospective	Examinee-centred, post hoc
Requires	Items with correct answers	Checklist plus global rating per candidate
Referencing	Criterion	Criterion

Borderline regression, which I covered in detail in One Hidden Error, inverts the problem: rather than asking experts to imagine a borderline candidate in the abstract, it asks examiners to *recognise* one in front of them. Each examiner records both a checklist score and a separate holistic global rating; regress one on the other; the height of the fitted line above "borderline" is the cut score for that station.

Two different instruments, two different methods, **both criterion-referenced**. That last point is what holds the architecture together. Neither method sets a pass rate in advance, and neither passes a fixed proportion of the cohort. They arrive at a criterion by different routes because the evidence they work from has a different shape.

10. A closing note that is a little too on-the-nose

While checking the literature for this post I found a paper published in *Medical Teacher* on 11 June 2026: *Secure AI-assisted Angoff standard-setting for single best answer questions: a non-inferiority validation study*.

The design: construct a description of a borderline student, build a secure offline item feature-extraction tool, generate AI-derived Angoff estimates, and test whether they fall within a prespecified margin of a human panel's.

Consider where that leaves this project. I am proposing to use a human panel to set the pass mark on an examination that tests whether clinicians can avoid deferring to AI output — and the literature is now asking whether that panel's judgement can itself be delegated to an AI.

I am not against it, and I want to be careful not to be reflexively against it, because that is exactly the under-trust failure mode the Unit 1.5 curve warns about. The Delegation question in the framework I have built on is not *should we use AI here* but *what is the cost if this is wrong and nobody notices?*

For standard setting, that cost is specific and large. A cut score sits behind every certificate, is rarely revisited, and is invisible in operation — precisely the profile of an error that propagates silently for years. So my answer for now is a boundary rather than a refusal:

AI-derived estimates may be admitted as a reality check, never as the panel. They can occupy the same slot the empirical p-values occupy in section 6 — a calibration input that judges see after round one, must consider, and may reject with their reasons recorded. What they cannot do is replace the human judgement about where the line goes, because that judgement is not a prediction about item difficulty. It is a decision about how much risk we are prepared to accept on a patient's behalf, and there is nobody to hold accountable for it except the people who made it.

That is the same rule the curriculum teaches clinicians about their own practice. It would be difficult to defend applying a laxer one to the exam.

Sources

- **The original** — Angoff, W.H. "Scales, norms and equivalent scores," in Thorndike, R.L. (ed.), *Educational Measurement*, 2nd edn, American Council on Education, 1971. The method appears in a footnote to Angoff's description of Tucker's yes/no procedure.
- **Systematic review and meta-analysis, 91 studies** — "Angoff methods in standard setting in health professional education: a systematic review and meta-analysis," *BMC Medical Education*, 2025;25:1727. doi:10.1186/s12909-025-08300-6 — the source for every pooled statistic quoted in section 7.
- **Judges and item difficulty** — Impara, J.C. and Plake, B.S. "Teachers' ability to estimate item difficulty: a test of the assumptions in the Angoff standard setting method," *Journal of Educational Measurement*, 1998. doi:10.1111/j.1745-3984.1998.tb00528.x
- **Anchor statements** — Angoff anchor statements: setting a flawed gold standard?

- **Practical standard-setting guidance** — AMEE Guide No. 18: Standard setting in student assessment (PDF), covering Angoff, Ebel, Hofstee and the compromise methods.
- **Yes/No as a substitute for percentage Angoff** — a simulation study on the Korean Medical Licensing Examination, *Journal of Educational Evaluation for Health Professions*, 2023.
- **AI-assisted Angoff** — Stephenson, E. et al. "Secure AI-assisted angoff standard-setting for single best answer questions: a non-inferiority validation study," *Medical Teacher*, published online 11 June 2026. doi:10.1080/0142159X.2026.2681212
- **Borderline regression, for the performance side** — Wood, T.J., Humphrey-Murto, S.M. and Norman, G.R. "Standard setting in a small scale OSCE," *Advances in Health Sciences Education*, 2006. doi:10.1007/s10459-005-7853-1
- **The automation-bias evidence behind the 40% Discernment weighting** — *Automation Bias in Large Language Model–Assisted Diagnostic Reasoning among Physicians Trained in AI Literacy*, NEJM AI, 2025

Source documents: the blueprint · Annex B — Competency Standards · Annex C — Level 1 Common Core · the Level 1 facilitator deck, whose final slide is the subject of this post.

Drafted with AI assistance. The worked panel is illustrative, but the judge cut scores, panel cut score and standard deviations are computed from the published estimate matrix rather than asserted — you can check every figure in the two tables above. The design decisions, and any errors, are mine.