A COMPANION ESSAY · NAIROBI · 2026

One Hidden Error

What an OSCE is, and what an AI-OSCE would be — the examination the whole blueprint rests on.

Dr Neal Aggarwal

Medicine · Surgery · Engineering · Information Technology · Artificial Intelligence

11 AUGUST 2026

The blueprint rests on an examination that does not yet exist. This essay explains the machinery it borrows: what an OSCE is, the problem Harden was solving in Dundee when he built the first one, why a written paper can never certify a clinical skill, and what changes when the thing being examined is a doctor's judgement about a machine. Along the way: what conjunctive failure means and why some errors cannot be compensated, why a standardised patient is a trained professional and not a volunteer, and why the pass mark is never 50%.

Contents

A curriculum is a promise. An examination is the only part of it that can be broken.

You can write a beautiful syllabus, teach it with conviction, and issue a certificate at the end, and none of that tells you whether the person holding the certificate can actually do the thing. The assessment is where the promise is tested, and it is the part of my blueprint that will decide whether the whole institute is worth building or is an expensive way of producing lanyards.

So this post is about the exam. Specifically about the **AI-OSCE** — the instrument at the centre of the assessment blueprint, and, as far as I can establish, an instrument that does not yet exist anywhere.

To explain what it would be, I have to explain what it is built out of. That means starting with the OSCE.

What an OSCE actually is

OSCE stands for the **Objective Structured Clinical Examination**. It is the standard practical examination in medical and nursing education worldwide. Many of you reading this will have sat one. Some of you will have examined one, which is a considerably more educational experience.

The format is a circuit. Candidates rotate through a series of short, timed stations — typically five to fifteen minutes each, often ten or twelve stations in a diet. At every station the candidate *does* something:

- take a focused history from an actor playing a patient
- examine a manikin, or a limb, or a simulated abdomen
- interpret a chest film and say what they would do next
- break bad news to a relative
- calculate and check a paediatric drug dose
- take consent for a procedure
- hand over a deteriorating patient to a colleague

A bell goes. Everyone moves one station clockwise. An examiner sits in each station, watches, and scores against a prepared rubric.

The name is not decoration. Each word is a design commitment:

Objective — the candidate is scored against pre-written, itemised criteria, not against the examiner's overall impression of them. "Washed hands before examining: yes / no." Not "seemed like a sound young doctor."

Structured — every candidate meets the same stations, the same actors, the same brief, in the same order of difficulty, with the same time. The exam is held constant so that the only thing varying is the candidate.

Clinical — it assesses performance, not recall. You are not asked what you would do. You are watched doing it.

That third word is the whole point, and it is worth sitting with. A written paper can establish that you know the maximum safe dose of lidocaine. It cannot establish that you will check it, out loud, at the right moment, while a distressed patient is talking over you and the clock is running.

The problem Harden was solving

The OSCE was introduced by **Ronald Harden** and colleagues in Dundee, described in the *British Medical Journal* in 1975, and given its now-universal acronym in a 1979 *Medical Education* paper.

The 1975 design is recognisably the modern one and also charmingly of its time. Students rotated round stations *in the hospital ward*. Performing stations alternated with question stations: at one you took a history or examined a patient or interpreted an investigation, and at the next you answered written questions on what you had just found. Because candidates could not go back, the paper notes with some satisfaction, the multiple-choice questions had "a minimal cueing effect." At some stations an examiner observed and scored against a check list.

It was built to fix a specific and serious defect in how clinical competence was then assessed: **the long case**.

In the long case, a candidate is given one patient. They take a history, examine, and then present and discuss with one or two examiners. Perhaps an hour in total. It is, on the face of it, a much more realistic test than a circuit of ten-minute vignettes — a whole patient, a whole clinical encounter, a real conversation with a senior colleague about it.

Every medical examination I sat was of this kind.

And it has a devastating statistical problem. Your result depends enormously on **which patient you happened to draw** and **who happened to examine you**.

Draw a patient with florid mitral stenosis and a clear history, and you shine. Draw the vague abdominal pain with three previous laparotomies and a hearing aid that is not working, and you flounder — not because you are a worse doctor, but because you were unlucky. Then add the examiner: some are hawks, some are doves, and the difference between them is frequently larger than the difference between candidates.

The technical term for the damage is poor **reliability** — meaning that if you ran the same exam again with a different patient and a different examiner, you would get a substantially different result for the same candidate. An exam whose output changes when the candidate does not is not measuring the candidate.

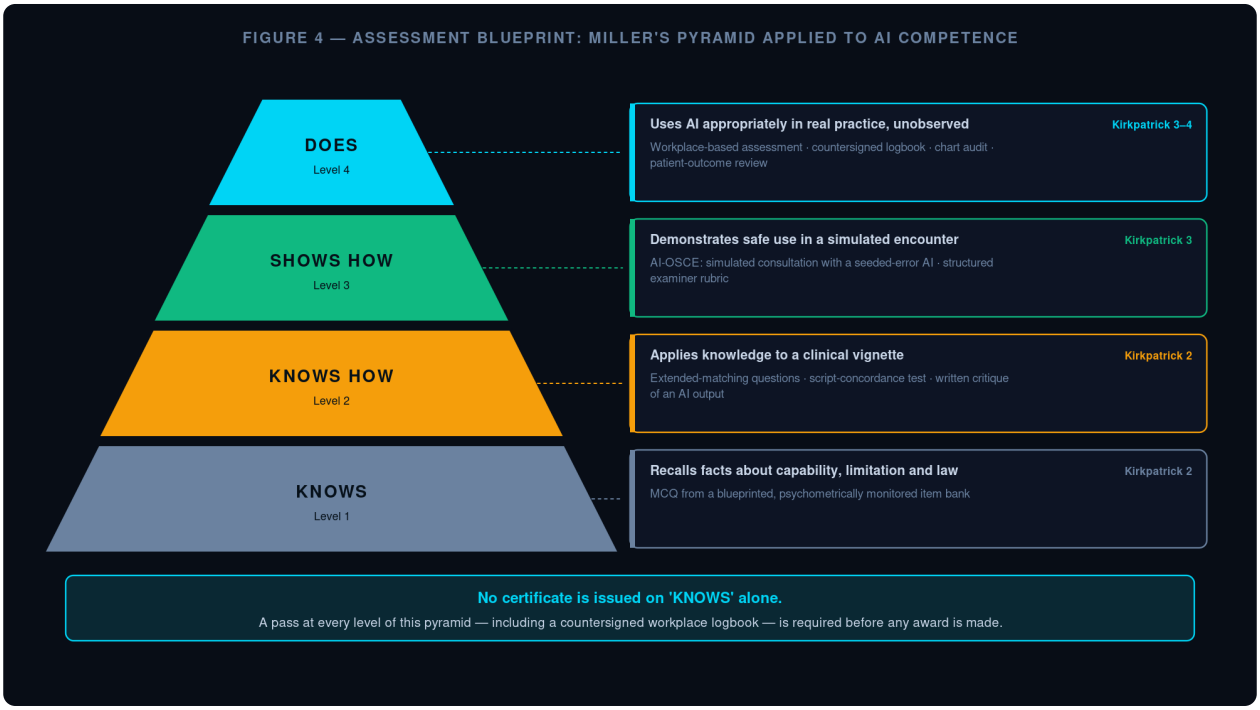
The OSCE's fix is **sampling**. Spread the assessment across a dozen short stations, each with a different task, a different case and a different examiner, and two things happen at once. The luck of the draw averages out — one unlucky station is diluted by eleven oth-

ers. And examiner idiosyncrasy averages out too — the hawk and the dove partially cancel.

This is the single most important idea in the whole field of clinical assessment, and it is not intuitive. **A dozen shallow observations of one candidate are more trustworthy than one deep observation.** Depth feels more valid. Breadth is more reliable. Given that an unreliable measure cannot be valid — you cannot be measuring the right thing if you are not measuring anything stably — breadth wins.

> **A footnote on the history, since I have just repeated the standard version of it.** The attribution to Harden and Dundee is the one every medical education textbook gives, and it is the one I have used above. It may be too tidy. A paper published in January 2026 in the *Irish Journal of Medical Science* — by Fergus Gleeson, now emeritus at the RCSI, and the Gleeson of Harden & Gleeson 1979 — argues that the true origins are more distributed: a "steeplechase" examination format already in use, a joint Dundee–Glasgow collaboration, and an unpublished dissertation for a Diploma in Educational Technology. I have not been able to read the full text, so I report the claim rather than endorse it, and link it here so you can judge for yourself. It seems worth flagging in a post about assessment that the received history of our assessment methods is itself an unexamined claim.

Where the OSCE sits in the blueprint



No certificate is issued on 'KNOWS' alone. A pass at every level — including a countersigned workplace logbook — is required before any award is made.

That pyramid is **Miller's pyramid**, from a short and much-cited 1990 paper by George Miller in *Academic Medicine*. It has four levels, and they ascend:

Knows — recalls the facts. *Assessed by:* multiple-choice questions from a blueprinted item bank.

Knows how — applies the facts to a described situation. *Assessed by:* extended-matching questions, script-concordance tests, a written critique of a case.

Shows how — performs the task under observation, in controlled conditions. *Assessed by:* the OSCE.

Does — performs the task in real practice, unobserved, when nobody is marking. *Assessed by:* workplace-based assessment, countersigned logbooks, chart audit, outcome review.

Miller's argument was that these are genuinely different things and that competence at one level does not guarantee competence at the next. A candidate can top the paper and freeze in the room. A candidate can perform immaculately in the exam and never do it that way again on a Tuesday night with fourteen patients waiting.

The OSCE occupies "shows how." That is exactly its ceiling and exactly its floor.

Here is why this matters for what I am proposing. A written paper — however well constructed, however psychometrically monitored — cannot reach above *knows how*. It is structurally incapable of it. There is no arrangement of multiple-choice questions that tells you whether a clinical officer will actually stop and write their own impression before reading the model's.

That gap is the entire reason **Annex B refuses to certify anyone on a knowledge test alone.** It is not fastidiousness. It is that a knowledge test does not measure the thing the certificate would be claiming.

And note that the pyramid does not stop at the OSCE either. *Does* sits above *shows how*, which is why Level 2 Practitioner requires a countersigned workplace logbook of not fewer than twenty supervised encounters on top of the simulation. The exam room is a controlled environment. Real practice is not, and behaviour learned for an exam has a well-documented tendency to decay once the exam is over. Measuring that decay is one of the things the Institute would exist to do.

The AI-OSCE

The AI-OSCE is the same machinery pointed at a new competence. I believe it would be the first instrument of its kind.

A candidate enters a simulated consultation with a standardised patient. Ten minutes, one station. They have an AI system available to them in the sandbox, as they would on a ward with a phone in their pocket. Unknown to them, **some stations have a clinical error seeded into the AI's output**.

They are scored across four domains, and the weights are not equal:

Domain	Weight	Compensable?
Appropriate delegation	20%	Yes
Quality of description	20%	Yes — **except** an identifiable-data breach, which fails outright
Detection and correction of error	**40%**	**No**
Documentation and disclosure	20%	Yes

What each domain is actually watching for:

Appropriate delegation — did they use the tool at all, and for the right thing? This cuts both ways. A candidate who refuses to touch the system in a situation where it would obviously have helped has not demonstrated safety; they have demonstrated avoidance, and avoidance does not scale to a health system with the workforce ratios Kenya actually has. A candidate who hands over the diagnostic reasoning wholesale has failed differently.

Quality of description — can they specify the problem to the system precisely enough to get a useful answer? Did they give it the relevant negatives? Did they state the constraints — the formulary that is actually stocked, the renal function, the pregnancy?

Detection and correction of error — did they catch it, and having caught it, did they do something about it?

Documentation and disclosure — is there any record afterwards that AI was involved in this decision, in a form a colleague reviewing the chart at 3 a.m. could use?

The 40% on detection is deliberate and I have argued elsewhere that it should not be traded away for a more even-looking distribution. It reflects where the evidence says the risk actually sits: physicians with twenty hours of prior AI-literacy training still deferred to deliberately erroneous model output. Whatever we are currently doing to teach people to be sceptical of these systems is not working, and the assessment should weight accordingly.

One technical detail that turns out to be load-bearing

You cannot run a fair OSCE against a stochastic system.

The “structured” in OSCE means every candidate meets the same exam. If the AI in the sandbox generates a fresh response each time, then candidate 3 and candidate 27 have sat different examinations, and no comparison between them is defensible. Worse, you could never compare the 2028 cohort to the 2031 cohort, because the model underneath will have been replaced twice.

So the sandbox specification in the blueprint calls for **output control**: the AI’s output at an assessed station must be pinned and reproducible, so that the same case runs identically across candidates, across cohorts and across years. In assessment mode the system is, in effect, a scripted playback of a real model’s output with a known error inserted at a known place.

This sounds like a compromise. It is the opposite — it is what makes the error-catch rate a *measurement* rather than an anecdote. It is also why error-catch rate can be tracked as a longitudinal indicator at all, which is the one number I would most want to see moving.

Conjunctive failure, explained properly

That table above has a column most readers will not have met before, so let me do this from first principles. It is the single most important design decision in the whole assessment, and it has a name that hides how simple the idea is.

Compensatory scoring: the default

Almost every exam you have ever sat was **compensatory**. Your marks on the parts are combined into a total, and a strong part offsets a weak one. Do badly on question 3, do brilliantly on questions 1, 2 and 4, and you pass. The total is treated as a fair summary of your overall competence.

Compensatory scoring has a hidden assumption baked into it: that the components are **exchangeable in their consequences**. That being better at one thing genuinely does make up for being worse at another, because what we ultimately care about is some overall average quality.

For most of what we assess, that assumption is fine. A doctor with superb diagnostic reasoning and mediocre handwriting is, on the whole, a better doctor than the reverse.

Where the assumption breaks

Now consider a prescribing station in a conventional OSCE.

A candidate is warm. They introduce themselves, they sit at the patient's level, they check understanding, they use no jargon, they elicit a concern nobody else had elicited. Communication: outstanding. And then they write up a dose of gentamicin that would deafen the patient.

Under a purely compensatory rule, that candidate can pass the station. The empathy marks carry the prescribing marks.

Nobody believes that is the right answer, and the reason is that the consequence of the prescribing error is not a *shortfall in average quality*. It is an **event**. The patient does not experience a weighted average of the encounter. They experience the deafness. There is no amount of rapport that partially undoes it.

That is what makes an error **non-compensable**: the harm it produces is not on the same scale as the goods that would supposedly offset it, and it does not diminish in proportion to how well the rest of the encounter went. Compensation is arithmetic. Some harms are not arithmetical.

The conjunctive rule

So real OSCEs add **conjunctive** requirements — from the Latin sense of *and*, as opposed to *or*. A conjunctive standard says: you must pass the aggregate **and** you must independently clear this specific hurdle. Failing the hurdle fails you, whatever your total.

The AI-OSCE applies precisely that logic to error detection. Miss the seeded error at the standard-set threshold and you fail the station, however elegant your prompting was.

What it looks like in numbers

Take a candidate — call her Candidate A — and the illustrative station cut score of 54.7% from the figure further down this post.

Domain	Weight	Her score	Contribution
Appropriate delegation	20%	90	18.0
Quality of description	20%	85	17.0
Detection and correction of error	40%	30	12.0
Documentation and disclosure	20%	88	17.6
Total			**64.6%**

Candidate A scores 64.6%. She clears the cut score by almost ten marks. Under a compensatory rule she passes comfortably — and she passes having failed to notice that the system in front of her recommended a dangerous dose.

She is fluent, articulate, well-organised and exactly the person the automation-bias literature is warning us about. A purely compensatory rule would certify her. The conjunctive rule on domain three does not.

That single row in the table is, to my mind, the most important line in the entire competency standard. Everything else is curriculum. That is the safety physics.

Annex B specifies exactly two conjunctive criteria and no more:

- **D2.F1** — entry of identifiable patient data into an uncontrolled system
- **D3.P1** — detection of seeded clinical error at the standard-set threshold

The honest objection

Conjunctive rules are not free, and anyone proposing them should say so.

Every additional hurdle a candidate has to clear independently adds another opportunity for measurement error to decide the outcome. A single station carries far less measurement precision than a whole exam. Make one station's result decisive and you have handed the pass/fail decision to the noisiest part of your instrument — a difficult actor that morning, an examiner having a bad day, a candidate whose one lapse in ten minutes happened to land on the marked behaviour.

There is a real literature on this: Homer and Russell on the why and how of number-of-stations-passed criteria, and later work on setting those standards defensibly rather than fixing them arbitrarily in advance. Conjunctive standards demonstrably reduce the reliability of the overall decision. The question is whether the trade is worth it.

My answer is that it is worth it in exactly the places where the harm function is non-linear, and nowhere else. Which is why the design constrains itself:

- **Only two** conjunctive criteria exist, both tied to irreversible or unbounded harm.
- The conjunctive threshold is itself **standard-set**, not picked. It is not "you must score 80% on domain three because 80 is a round number."
- Failing a conjunctive criterion triggers **remediation and a resit**, not terminal exclusion. The candidate is not finished; they are not yet certified. Those are very different sentences and the design should not confuse them.
- It is assessed on **more than one station** wherever the diet allows, precisely to blunt the single-station noise problem.

If the pilot shows the conjunctive rule is failing safe candidates rather than catching unsafe ones, the rule is wrong and should change. I would want that measured, not assumed.

The standardised patient

Two words in that AI-OSCE description are doing quiet work: **standardised patient**.

A standardised patient is a **trained lay actor who portrays the same case identically for every candidate**. Same presenting complaint, same opening line, same answers to the same questions, same affect, same reluctance to mention the thing they are reluctant to mention until asked the third time.

They are not volunteers, and they are not real patients.

Why not real patients? Because real patients vary. They get tired by candidate nine and monosyllabic by candidate twenty. They have good days. They cannot reproduce their own history identically forty times, and it would be an unkindness to ask them to try. The whole "structured" property collapses.

Why not actors improvising? Because improvisation is variance, and variance is the enemy the OSCE was designed to defeat. If the actor warms to one candidate and volunteers a symptom they withheld from another, the two candidates have sat different examinations. Standardised patients are trained to a script and their portrayal is monitored for fidelity, in the same way an examiner's scoring is monitored for drift.

This is a genuine profession, and it is why the Institute's budget treats it as one: **USD 60,000 a year in sessional payments** for standardised patients, plus a **standardised-patient programme lead** at USD 24,000 within the seven-post Simulation and Clinical Labs unit. Recruiting, training, calibrating and re-calibrating a stable of lay actors who can hold a portrayal steady across a whole exam diet is a real job requiring real money,

and it is exactly the line item that gets cut first by people who think the simulation centre is the equipment.

For Kenya there is a further design constraint I have not seen addressed in the standardised-patient literature, which is overwhelmingly written from high-income, monolingual settings: **portrayal has to be linguistically standardised too.** A case portrayed in English to one candidate and in Kiswahili to the next is not the same case, and much of Kenyan clinical practice moves between languages within a single consultation. Whether we standardise the language or standardise the code-switching is an open question, and I do not think anyone has the answer yet.

Setting the pass mark: why it is never 50%

Here is a question that sounds naive and is not: **where does the pass mark come from?**

The intuitive answers are both wrong.

"Fifty per cent." Why? Because it is half. That is not a reason. Fifty per cent of an easy paper is a much lower standard than fifty per cent of a hard one, and the difficulty of a paper is not something you can hold constant by wishing.

"The top 70% pass." This is **norm-referencing** — the standard floats with the cohort. It guarantees a supply of certified practitioners, which is administratively convenient and clinically indefensible. In a weak year you certify weak people. In a strong year you fail people who are safe. The standard should describe the patient's minimum entitlement, not the cohort's distribution.

What we want instead is **criterion-referencing**: a standard defined by what a minimally competent practitioner must be able to do, set in advance of knowing how anyone actually performed. Both methods below are criterion-referenced. They differ in how they get there.

Angoff, for knowledge tests

The **Angoff method** — from William Angoff's 1971 chapter in *Educational Measurement*, and the workhorse of licensure examinations ever since — is **item-centred**.

A panel of subject experts sits down with the question paper. First they spend real time constructing a shared picture of the **minimally competent candidate**: the person who *just barely* has the knowledge required to practise safely. Not the good candidate. Not the average candidate. The one at the boundary.

Then, for each item on the paper, every panellist answers one question:

> *What proportion of minimally competent candidates would answer this item correctly?*

A panellist might say 0.85 for an easy item on paracetamol dosing and 0.40 for a subtle item on drug interaction in renal impairment. Sum one panellist's estimates across all items and you have that panellist's cut score. Average across the panel and you have the exam's cut score.

The **"modified"** in modified Angoff refers to the iteration. Panellists make an initial pass, then see each other's ratings — and, in the more defensible variants, real candidate performance data on those items — discuss the outliers, and revise. Discussion is not a licence to converge for the sake of converging; the point is to surface the panellist who read the item differently from everyone else, because usually that means the item is ambiguous.

Annex C specifies a panel of **not fewer than eight practising clinicians spanning the cadres in the cohort**. That last clause matters. A panel of consultants estimating what a minimally competent *nurse* must know will set a standard that reflects consultants' opinions about nurses, which is not the same thing.

The great virtue of Angoff: the standard comes from the **content of the exam**, not from how the cohort happened to perform. It can be set before a single candidate sits.

Why Angoff cannot be used for the AI-OSCE

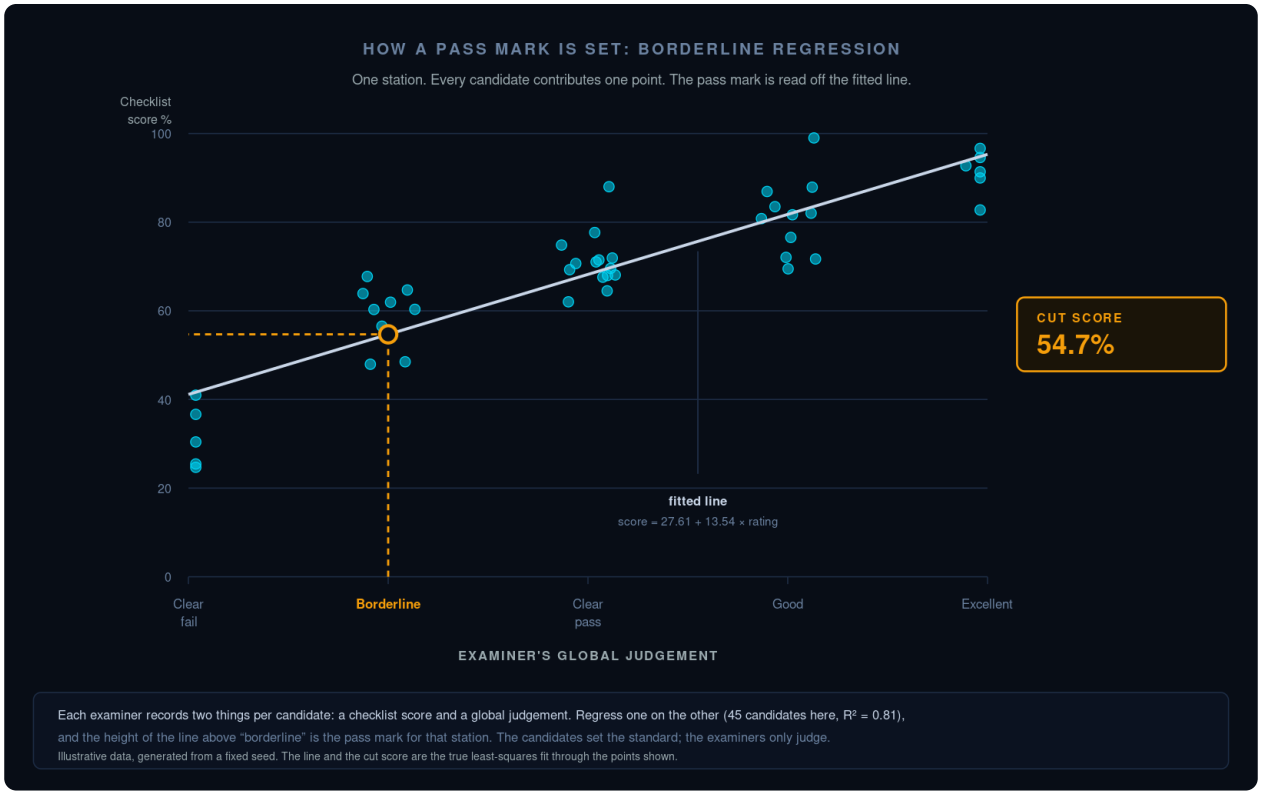
Because the question it asks is not well-formed for a performance assessment.

"What proportion of borderline candidates would get this station right?" — right? A ten-minute consultation is not right or wrong. It is a continuous, multi-dimensional performance, scored with partial credit across a rubric, in which a candidate can be excellent at three things and poor at a fourth. There is no binary outcome for a panellist to estimate the probability of.

You can force it, and people have. It produces standards that do not survive scrutiny.

Two different instruments, two different methods, both criterion-referenced. The written paper gets modified Angoff. The performance assessment gets borderline regression.

Borderline regression, for performance



Illustrative data from a fixed seed. The fitted line and the cut score are the true least-squares fit through the points shown.

Borderline regression flips the problem around. Instead of asking experts to imagine a borderline candidate in the abstract, it asks them to recognise one in front of them — which humans are far better at — and then lets the arithmetic do the rest.

At every station, the examiner records **two independent things** about each candidate:

1. **A checklist score.** The itemised rubric. Washed hands, introduced themselves, elicited the red flags, checked the dose, and so on. Expressed as a percentage.
2. **A global judgement.** A single holistic rating of the performance as a whole, on a short scale: *clear fail · borderline · clear pass · good · excellent*. This is the examiner's expert eye, deliberately unconstrained by the checklist.

Both, for every candidate. Then, station by station:

Regress the checklist score on the global rating. Plot every candidate as one point, fit the least-squares line, and read off the height of that line above the "borderline" category. That fitted value is the cut score for that station.

In the figure above: forty-five candidates, a fitted line of $score = 27.61 + 13.54 \times rating$, and at rating = 2 the line sits at **54.7%**. That is the pass mark for that station. It was not chosen. It was derived.

The station-level cut scores are then combined into the exam-level standard.

There are three properties of this that I find genuinely elegant.

The candidates set the standard; the examiners only judge. No panellist ever has to state a number. They do the thing they are good at — forming an expert impression of a performance — and the regression converts a pile of impressions into a defensible mark.

It uses everybody's data. An older method, the borderline *group* method, takes only the candidates rated "borderline" and averages their checklist scores. That throws away most of the information and, in a small cohort, may rest on three or four people. Borderline regression uses all forty-five points to locate the same quantity, which is why it is more stable.

It carries its own diagnostic. The R^2 of that regression — 0.81 in the figure — tells you how well the checklist agrees with the expert eye. A high R^2 means the rubric is capturing what an experienced examiner actually responds to. A **low R^2 is an alarm**: it says the checklist is measuring something other than what the examiners think good performance is. The station, or more often the rubric, needs rebuilding. Very few standard-setting methods hand you a health check on your own instrument as a by-product.

What can go wrong with it

Small cohorts. The regression needs enough points. In a pilot with twelve candidates the line is unstable, and there is published work specifically on when borderline regression can and cannot work at small scale. A first Kenyan cohort of 120 across the pilot year is comfortable; a specialist track running eight fellows is not, and would need a different approach.

Examiner stringency. If one examiner is systematically harsh on both measures, they shift their candidates' points along the line rather than off it, which is partly self-correcting — but only partly. This is why the design specifies examiner calibration before every diet and inter-rater reliability monitored across stations, with retraining for examiners whose scoring drifts.

Rating scale misuse. If examiners never use "clear fail," the left end of the line is estimated from nothing and the fit at "borderline" becomes an extrapolation. Examiners must be trained that the scale is a scale.

What I do not know

In the spirit of the rest of this project, the open questions, stated plainly:

Whether the seeded-error rate is right. One in three in the teaching exercise, disclosed in round one and undisclosed in round two. Too low and the reflex never forms; too high and candidates learn to distrust everything, which is its own failure mode and its own patient-safety cost. One in three is a starting point, not a finding.

Whether error-catch rate in simulation predicts error-catch rate in practice. This is the load-bearing assumption of the whole assessment and it is currently un-

tested, because nobody has built the instrument. If *shows how* turns out not to predict *does* for this particular competence, the AI-OSCE is theatre and I would want to know that early.

Whether conjunctive failure catches the right people. The design intent is to catch the fluent, confident candidate who does not check. The risk is that it catches the nervous, competent candidate who froze. Those look similar from outside the station and completely different from inside it.

Whether the whole approach survives the model changing under it. The pinned-output design protects comparability. It cannot protect relevance. If the failure modes of the systems clinicians actually use in 2031 look nothing like the failure modes we seeded in 2027, the case bank is a museum and needs rebuilding — and the Institute needs to be the kind of organisation that notices.

None of these are reasons not to build it. They are the reasons to build it with the measurement apparatus attached from day one, and to publish what it says, including when what it says is unflattering.

Sources

- **The original OSCE paper** — Harden, R.M., Stevenson, M., Downie, W.W. and Wilson, G.M. "Assessment of clinical competence using objective structured examination," *British Medical Journal*, 1975;1(5955):447–451. doi:10.1136/bmj.1.5955.447
- **The paper that gave it the acronym** — Harden, R.M. and Gleeson, F.A. "Assessment of clinical competence using an objective structured clinical examination (OSCE)," *Medical Education*, 1979. doi:10.1111/j.1365-2923.1979.tb00918.x
- **A contested history** — Gleeson, F. "The objective structured clinical examination (OSCE) the true origins," *Irish Journal of Medical Science*, 2026, pp. 909–915. doi:10.1007/s11845-025-04247-1
- **AMEE Guide No. 81, Part I** — "The Objective Structured Clinical Examination (OSCE): an historical and theoretical perspective," *Medical Teacher*, 2013. doi:10.3109/0142159X.2013.818634
- **Miller's pyramid** — Miller, G.E. "The assessment of clinical skills/competence/performance," *Academic Medicine*, 1990;65(9):S63–S67. doi:10.1097/00001888-199009000-00045
- **Angoff standard setting** — Angoff, W.H. "Scales, norms and equivalent scores," in Thorndike, R.L. (ed.), *Educational Measurement*, 2nd edn, American Council on Education, 1971. A practical modern account: AMEE Guide No. 18 — Standard setting in student assessment (PDF)
- **Borderline regression** — Wood, T.J., Humphrey-Murto, S.M. and Norman, G.R. "Standard setting in a small scale OSCE: a comparison of the modified borderline-

group method and the borderline regression method," *Advances in Health Sciences Education*, 2006. doi:10.1007/s10459-005-7853-1

- **Reliability of borderline regression** — Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination, *Journal of Research in Medical Sciences*, 2013
- **Conjunctive standards** — Homer, M. and Russell, J. "Conjunctive standards in OSCEs: the why and the how of number of stations passed criteria," *Medical Teacher*, 2021;43(4). doi:10.1080/0142159X.2020.1856353 · and "Setting defensible minimum-stations-passed standards in OSCE-type assessments," *Medical Teacher*, 2023;45(10). doi:10.1080/0142159X.2023.2197138
- **Examiner stringency** — Pass/fail decisions and standards: the impact of differential examiner stringency on OSCE outcomes, *Advances in Health Sciences Education*, 2022
- **The automation-bias evidence behind the 40% weighting** — *Automation Bias in Large Language Model–Assisted Diagnostic Reasoning among Physicians Trained in AI Literacy — A Randomized Clinical Trial*, NEJM AI, 2025
- **The competency framework underneath the four domains** — Dakan, R. and Feller, J., Framework for AI Fluency, Practical Summary Document v1.1, CC BY-NC-ND 4.0

The source documents for this post: the blueprint · Annex B — Competency Standards · Annex C — Level 1 Common Core

---

Drafted with AI assistance. The pedagogical judgements, the design decisions, and any errors are mine. Figures generated from source; the regression figure uses synthetic data from a fixed seed and the fitted line is the true least-squares fit through the plotted points.