Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When the grade disappears: A natural experiment examining clinical- skills performance after removal of grade stakes from an early objective structured clinical examination

Tetiana Hliebova

THliebova@rossu.edu

Ross University School of Medicine

Rakesh Calton

Ross University School of Medicine

Research Article

Keywords: Educational measurement, clinical competence, formative feedback, motivation, objective structured clinical examination

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11405841/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Early formative (ungraded) objective structured clinical examinations (OSCEs) are increasingly used to reduce novice-learner stress and promote feedback. However, measurement theory holds that scores on low-stakes assessments reflect not only ability but also test-taking effort, a concern rarely quantified in medical education. We examined the effect of removing grade stakes from an early clinical-skills OSCE and whether performance recovered when stakes returned.

Methods

We conducted an observational natural experiment at an offshore medical school. One first-semester cohort (Summer 2026, n = 241) completed the week-6 OSCE as an ungraded formative assessment; three preceding cohorts had completed the same examination summatively. All cohorts subsequently completed a graded week-12 OSCE. Both examinations included an interview station assessing aligned history-taking and communication domains. Blood-pressure measurement and vital-sign interpretation were assessed using identical criteria at both examinations and provided a content-controlled within-student anchor. Analyses used Welch's t-tests with Mann–Whitney U confirmation, Cohen's d, distributional diagnostics, and paired within-student comparisons.

Results

Under formative conditions, performance was lower than in each of the three preceding summative cohorts on both the interview station (7.9–10.3 percentage points) and physical-examination station (9.8–13.7 points) (Welch t, all $p < .001$; Cohen's d 0.59–0.86). The decline was distribution-wide rather than driven by outliers and was evident across academic-performance levels: even the strongest written-examination quartile scored below prior cohorts. When stakes returned at the graded OSCE, the same students showed substantial improvement: interview performance increased by a mean of 9.6 points within students (paired $d = 0.91$, $p < .001$), and the identical blood-pressure/vital-signs anchor rose 5.7 points (paired $d = 0.36$, $p < .001$). Compared with historical week-12 cohorts, recovery was near-complete (residual deficits 2.0–5.6 points; small effects).

Conclusions

Transitioning an early OSCE from summative to formative assessment was associated with a large, cohort-wide decline in clinical-skills performance, followed by substantial within-student recovery when stakes returned. The magnitude, breadth, and reversibility of these findings are consistent with assessment stakes – and potentially test-taking effort – being an important contributor to observed

OSCE performance. Formative OSCE scores should therefore be interpreted cautiously when used to infer clinical competence or identify students for remediation.

Trial registration:

Not applicable.

Background

The objective structured clinical examination (OSCE) is a mainstay of clinical-skills assessment in undergraduate medical education [1, 2]. Contemporary assessment theory distinguishes assessment of learning from assessment for learning, and programmatic assessment frameworks argue that any single method should be judged not only on reliability and validity but on its educational impact [3, 4]. Within this view, converting an early OSCE from summative to formative is intended to lower stakes, reduce anxiety, and create space for feedback and self-regulated learning [5].

A substantial body of measurement research, however, shows that scores on low-stakes assessments reflect not only ability but also test-taking effort. In experimental studies where motivation was manipulated, unmotivated examinees underperformed motivated ones by up to 0.6 standard deviations, with corresponding threats to score validity [6]. Expectancy-value theory explains this: when a task carries no consequence, its subjective value and therefore the effort invested declines [7, 8]. Cilliers and colleagues demonstrated specifically that the consequences attached to an assessment shape how, and how hard, medical students engage with it [9]. Consequential validity, which is the impact of an assessment's use on behaviour, is therefore central to interpreting a formative OSCE.

Evidence on whether early formative OSCEs improve later summative performance is mixed and, importantly, station-specific. In a large cohort, formative OSCEs improved subsequent summative performance only for identical stations and did not raise overall pass rates [10]; a randomised trial similarly found a single formative OSCE did not change summative pass rates [11]. These findings motivate a design anchored on identically repeated skills.

Despite the strength of the theoretical argument, the behavioural consequences of removing stakes from an early OSCE have rarely been quantified directly in medical education, and the question of whether any such effect is reversible, a strong test of the effort interpretation, has not, to our knowledge, been examined. An institutional change at our school created a natural experiment permitting exactly this test. Because many medical schools are moving early OSCEs to formative formats, the answer bears directly on how formative OSCE scores are interpreted beyond our institution. We asked: (1) Does removing grade stakes from an early OSCE change performance, and by how much? (2) Is any change distribution-wide or driven by a subset of students? (3) Is it explained by cohort ability? (4) Does performance recover when stakes return at a subsequent graded OSCE?

Methods

Design and setting

We conducted an observational, longitudinal natural experiment at an offshore medical school delivering a US-style four-year MD curriculum. Reporting follows the STROBE guideline for observational studies (completed checklist in Appendix 1). First-semester students complete two sequential modules within the Fundamentals and Foundations of Medicine course: Fundamentals and Foundations of Medicine 1 (FFM1) and Fundamentals and Foundations of Medicine 2 (FFM2). Each module includes a clinical-skills OSCE, administered at approximately week 6 (FFM1) and week 12 (FFM2) of the semester. Each OSCE comprises an interview station assessing history-taking and communication skills and a physical examination (PE) station, delivered with standardized patients and faculty examiners using analytic checklists and global ratings under timed conditions. The FFM1 OSCE used the same station format and assessment checklists for all four cohorts.

Historically, both the FFM1 and FFM2 OSCEs were summative (grade-bearing). For the Summer 2026 cohort, the FFM1 OSCE was converted to an ungraded formative assessment, while the FFM2 remained summative. This institutional change created a within-cohort transition from an ungraded formative OSCE (FFM1) to a graded summative OSCE (FFM2), with preceding cohorts that had completed FFM1 under summative conditions providing historical comparators.

Cohorts and comparisons

The primary (index) cohort matriculated in Summer 2026 (n = 241 across stations). Of these, 200 students had linked scores at both FFM1 and FFM2 and were included in within-student analyses; the remaining 41 did not have linked scores at both examinations and were excluded from those analyses. Three preceding cohorts – Summer 2025 (n = 154), Fall 2025 (n = 333), and Spring 2026 (n = 226) – completed the FFM1 OSCE under summative conditions and served as historical FFM1 comparators. For benchmarking performance at the subsequent summative FFM2 OSCE, historical cohorts with available skill-level FFM2 data (Spring 2025, Summer 2025, and Spring 2026) served as comparators. Cohorts for which only aggregate FFM2 scores were available were excluded from skill-level comparisons.

The identically repeated anchor

The interview stations at FFM1 and FFM2 assessed the same core domains of history-taking and communication using aligned assessment criteria, although the clinical cases differed. The PE content differed between the two OSCEs, with general examination skills assessed at FFM1 and neurological examination skills at FFM2. However, blood-pressure measurement and vital-signs interpretation were assessed at both OSCEs using identical checklists. These repeated skills provided a content-controlled, within-student anchor for examining change in performance between the formative FFM1 and

summative FFM2 OSCEs. Because the assessed skills and scoring criteria were held constant, changes in anchor performance were not attributable to differences in examination content, although effects of additional learning, practice, and skill maturation between assessments could not be separated from the change in assessment stakes. Blood-pressure measurement was included in FFM2 only for the index cohort and was therefore examined only in within-student analyses; historical FFM2 benchmarking of PE performance used vital-signs interpretation, which was assessed across cohorts.

Data and measures

De-identified station-level and total OSCE scores, expressed as percentages, were extracted for all cohorts. For the index cohort, scores from two interim written assessments and a cumulative final examination were also available and were used as an independent indicator of overall academic performance during the semester. These scores were used to examine whether patterns in formative OSCE performance differed across levels of academic performance. Records were linked across FFM1 and FFM2 using a temporary internal identifier, which was replaced with a coded study identifier before analysis; no directly identifiable information was included in the analytic dataset. Non-student administrative and summary rows were excluded before analysis.

Statistical analysis

Between-cohort differences were tested with Welch's t-test (not assuming equal variances), with the Mann-Whitney U test as a rank-based, outlier-robust confirmation; effect sizes are reported as Cohen's d. Ninety-five percent confidence intervals (CIs) are reported for FFM1 between-cohort comparisons and for within-student changes; CIs for mean differences were derived from the corresponding t-statistics and standard errors, and those for effect sizes from the standard large-sample approximation. To distinguish a whole-cohort shift from outlier effects, we compared mean and median movement, identified outliers using Tukey's fence (1.5 × IQR), recomputed results with outliers excluded and using 10% trimmed means, and calculated the proportion of students below prior-cohort reference points. To test whether ability explained the decline, we stratified the index cohort into written-examination quartiles and compared each quartile's OSCE performance with historical cohort means. Recovery was assessed by paired t-tests (with Wilcoxon signed-rank confirmation) on within-student FFM1→FFM2 change for the interview station and the anchor, and by benchmarking the index cohort's FFM2 performance against historical FFM2 cohorts. Within-student analyses used complete cases (students with scores at both examinations); no imputation was performed. Analyses were conducted in Python (SciPy library) and R; $\alpha = .05$, two-sided. Conclusions were unchanged after Holm correction for multiple comparisons. No a priori sample size calculation was performed, as the sample was determined by cohort enrolment.

Use of artificial intelligence tools

Claude (Anthropic) was used to assist with language editing, reference formatting, and the calculation of confidence intervals from the reported test statistics. The authors reviewed all output and take full responsibility for the content of the manuscript.

Ethics

This retrospective analysis of de-identified educational records was reviewed by the Ross University School of Medicine Research Committee and IRB and determined to be exempt (normal educational practice and secondary analysis of de-identified data; reference: RUSMIRB exempt #5-IRB2026_5). Analyses used de-identified individual-level records linked by a coded study identifier; no individual participant is identifiable in the reported results, and the requirement for individual consent was waived. All study methods were carried out in accordance with relevant guidelines and regulations, including the ethical principles of the Declaration of Helsinki.

Results

Performance was substantially lower under formative conditions

Under formative conditions, the index cohort scored markedly below every prior summative cohort on both stations (Table 1). On the interview station, the formative mean was 71.2% versus 79.0–81.5% for prior cohorts (differences of − 7.9 to − 10.3 points; Welch t, all $p < .001$; Cohen's d 0.62–0.86). On the PE station, the formative mean was 71.1% versus 80.9–84.7% (differences of − 9.8 to − 13.7 points; all $p < .001$; d 0.59–0.86). The Mann–Whitney U test confirmed every comparison (all $p < .001$).

Table 1
Formative FFM1 performance versus prior summative cohorts, by station

Station	Comparator (n)	Mean %	Δ, points (95% CI)	Welch p	d (95% CI)
Interview	Formative (241)	71.2	—	—	—
	Summer 2025 (154)	79.0	−7.9 (− 10.5 to − 5.3)	< .001	0.62 (0.41 to 0.83)
	Fall 2025 (333)	81.5	−10.3 (− 12.3 to − 8.3)	< .001	0.86 (0.69 to 1.03)
	Spring 2026 (226)	80.6	−9.5 (− 11.7 to − 7.3)	< .001	0.78 (0.59 to 0.97)
Physical examination	Formative (241)	71.1	—	—	—
	Summer 2025 (154)	84.7	−13.7 (− 16.9 to − 10.5)	< .001	0.86 (0.65 to 1.07)
	Fall 2025 (333)	82.4	−11.3 (− 13.9 to − 8.7)	< .001	0.71 (0.54 to 0.88)
	Spring 2026 (226)	80.9	−9.8 (− 12.8 to − 6.8)	< .001	0.59 (0.40 to 0.78)

Differences (Δ) are index cohort minus comparator. All comparisons by Welch's t-test with Mann–Whitney U confirmation (all p < .001). Cohen's d is reported as an absolute value. CI, confidence interval.

The decline was distribution-wide, not driven by outliers

The lower performance observed under formative conditions reflected a broad downward shift in the score distribution rather than being driven by a small number of very low scores. Medians declined in parallel with means. Low-score outliers, defined using Tukey's fence, comprised approximately 2% of the index cohort, and both exclusion of these observations and analysis using 10% trimmed means produced estimates like the primary analysis, with a difference of approximately 9 percentage points remaining (p < .001). Approximately four in five students in the formative cohort scored below the prior-cohort median, and approximately half scored below the prior-cohort 25th percentile, further supporting a broad distributional shift rather than an effect driven primarily by a small number of low-scoring students.

The decline was evident regardless of written-examination ability

Within the index cohort, OSCE performance increased across quartiles of written-examination performance, with a moderate positive correlation between written-examination and OSCE scores (Pearson r = 0.39). However, lower formative OSCE performance relative to historical summative cohorts was observed across all written-examination quartiles, including the highest-performing quartile (mean

written-examination score, 84%), whose mean OSCE performance remained at or below the overall means of the historical cohorts. Thus, the observed formative OSCE performance pattern was not confined to students with lower academic performance within the index cohort.

Performance improved when stakes returned

At the graded FFM2 examination, the same students improved substantially (Fig. 1, Table 2). The interview cohort mean rose from 71.2% at the formative FFM1 to 82.7% at the summative FFM2; among the 200 students with scores at both examinations, the mean within-student increase was 9.6 points (95% CI 8.1 to 11.1; paired $t = 12.9$, $p < .001$; $dz = 0.91$). The improvement was largest in the domains where formative scores had been lowest (interview process, history of presenting illness), although regression to the mean may contribute to this pattern.

On the content-controlled anchor (blood-pressure measurement and vital-signs interpretation), where examination content was held identical, the same students improved from 74.1% to 79.8%—a within-student increase of 5.7 points (95% CI 3.5 to 7.9; paired $t = 5.1$, $p < .001$; $dz = 0.36$), with 62% of students improving. Because content was constant, this change cannot be explained by differences in examination material, although the return of stakes cannot be separated from additional learning and skill maturation.

Recovery was near-complete but not total

Benchmarked against historical FFM2 cohorts sitting the same graded examination (Fig. 2), the index cohort remained slightly below prior cohorts on the shared, identically assessed skills: interview − 2.0 points ($p = .003$), vital-signs interpretation − 5.6 points ($p < .001$), and cranial-nerve examination − 4.5 points ($p < .001$; $d = 0.34$). Effect sizes were small. Thus, although the return of stakes reversed most of the formative deficit, the index cohort closed most – but not all – of the gap to prior cohorts.

Table 2
Within-student change from formative FFM1 to summative FFM2 (index cohort) and benchmark against historical FFM2 cohorts

Measure	FFM1%	FFM2%	Δ, points (95% CI); p	Effect size (95% CI)
Interview (within-student, n = 200)	71.2	82.7	+ 9.6 (8.1 to 11.1); p < .001	dz 0.91 (0.75 to 1.07)
Anchor: BP + vital signs (within-student, n = 200)	74.1	79.8	+ 5.7 (3.5 to 7.9); p < .001	dz 0.36 (0.22 to 0.50)
Interview vs historical FFM2	—	82.7	−2.0; p = .003	Small
Vital signs vs historical FFM2	—	83.9	−5.6; p < .001	Small
Cranial nerves vs historical FFM2	—	83.9	−4.5; p < .001	d 0.34

Interview FFM1 and FFM2 values are cohort means; Δ for within-student rows is the mean paired change among students with scores at both examinations. Benchmark Δ values are index cohort minus historical FFM2 cohorts. dz, paired effect size; BP, blood pressure; CI, confidence interval.

Discussion

Transitioning an early clinical-skills OSCE from summative to formative assessment was associated with a substantial, broad-based decline in performance relative to three preceding summative cohorts. The decline was not driven by a small number of low-scoring students and was observed across levels of academic performance within the index cohort. Importantly, longitudinal within-student analyses demonstrated substantial improvement when the same students subsequently completed a graded OSCE, including on the content-controlled repeated-skill anchor. Taken together, the magnitude, breadth, and subsequent recovery of performance are consistent with assessment stakes and potentially test-taking effort as contributors to observed OSCE performance, in keeping with research on motivation and low-stakes assessment [6, 9]. Because motivation and effort were not directly measured, however, this proposed mechanism remains to be tested prospectively.

The pattern of recovery varied across clinical-skills domains, with greater improvement in interview performance and more modest improvement in the repeated blood-pressure/vital-signs anchor. The anchor strengthens the longitudinal comparison because the assessed skills and scoring criteria were held constant across FFM1 and FFM2, reducing differences in examination content as an explanation for the observed change. Nevertheless, additional learning, practice, maturation, and increasing familiarity with the OSCE format may also have contributed to improvement. Examiners may also have scored differently when they knew the OSCE was ungraded, for example by applying stricter or more feedback-oriented standards, which could have lowered formative scores independently of student effort. Similarly, the small residual differences between the index cohort and historical FFM2 cohorts may reflect effects of the earlier formative assessment, cohort-level variation, or other unmeasured factors and cannot be distinguished in the present study.

These findings extend previous research examining whether formative OSCEs influence subsequent summative performance [10, 11] by highlighting an additional measurement consideration: performance observed during an ungraded formative OSCE may reflect not only clinical-skills proficiency but also the conditions under which students are asked to demonstrate that proficiency. This distinction is particularly relevant when formative OSCE scores are used diagnostically or to guide educational interventions.

The findings therefore have practical implications for formative assessment design. Performance on an ungraded formative OSCE should be interpreted cautiously when used as an indicator of clinical competence or to identify students for remediation and may be best considered alongside other sources of performance evidence. At the same time, formative OSCEs may benefit from strategies that maintain task value and meaningful learner engagement without reintroducing high-stakes grading. Structured

faculty-supported reflection, criterion-referenced self-assessment, video review, and facilitated debriefing are potential approaches supported by previous work [12–15].

Strengths and limitations

This study has several strengths, including a naturally occurring institutional change in assessment conditions, a large multi-cohort sample, and the combination of historical between-cohort and longitudinal within-student comparisons. The content-controlled repeated-skill anchor strengthened the longitudinal analysis by reducing differences in examination content as an explanation for changes in performance. Distributional and academic-performance analyses provided additional evidence that the observed decline was broad-based rather than driven primarily by a small subgroup of lower-performing students.

Several limitations warrant consideration. Most importantly, test-taking effort and motivation were not measured directly; therefore, reduced effort remains a plausible interpretation of the observed performance pattern rather than an established mechanism. The observational, single-institution design and use of historical comparators limit causal inference and generalizability and cannot exclude unmeasured differences between cohorts. We did not examine whether examiner and standardized-patient assignments were consistent across cohorts, and checklist reliability (internal consistency and inter-rater agreement) was not assessed; differences in examiner behaviour, including any change in scoring when the OSCE became formative, therefore cannot be excluded. Written-examination scores were available only for the index cohort, so academic performance could not be compared directly across cohorts; the quartile analysis compared index-cohort quartiles with historical whole-cohort OSCE means and therefore cannot exclude differences in cohort ability. Historical FFM2 data lacked blood-pressure assessment and complete skill-level data for some cohorts, limiting aspects of the benchmark comparison. In addition, physical-examination content differed between FFM1 and FFM2 outside the repeated-skill anchor, restricting direct longitudinal comparison of non-anchor PE performance. Finally, the change in assessment stakes is inseparable from additional learning, practice, maturation, and increasing familiarity with the OSCE format between assessments. Regression to the mean may also have contributed to the within-student gains, particularly among students with the lowest formative scores.

A prospective multi-cohort study incorporating a validated measure of test-taking effort (the Student Opinion Scale [16]) and structured observations of learner engagement at both OSCEs is underway to examine the proposed mechanism more directly.

Conclusions

Transitioning an early clinical-skills OSCE from summative to formative assessment was associated with a substantial, broad-based decline in observed performance that was followed by marked improvement when students subsequently completed a graded OSCE. The magnitude, breadth, and subsequent

recovery of performance are consistent with assessment stakes and potentially test-taking effort as contributors to observed OSCE performance. These findings suggest that scores from ungraded formative OSCEs should be interpreted cautiously as indicators of clinical competence and that formative assessment designs should incorporate strategies that support meaningful learner engagement while preserving opportunities for feedback and learning.

Abbreviations

BP
blood pressure
CI
confidence interval
FFM1
Fundamentals and Foundations of Medicine 1
FFM2
Fundamentals and Foundations of Medicine 2
IQR
interquartile range
OSCE
objective structured clinical examination
PE
physical examination
STROBE
Strengthening the Reporting of Observational Studies in Epidemiology.

Declarations

Ethics approval and consent to participate

This study was reviewed by the Ross University School of Medicine Research Committee and IRB and determined to be exempt (RUSMIRB exempt #5-IRB2026_5); the requirement for individual consent was waived for this retrospective analysis of de-identified educational data. All study methods were carried out in accordance with relevant guidelines and regulations, including the ethical principles of the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

TH and RC designed and delivered the assessments under study. The authors declare that they have no financial competing interests.

Funding

This study received no funding.

Author Contribution

Tetiana Hliebova conceived the study, contributed to the study design and interpretation of the findings, and drafted the manuscript. Rakesh Calton led the analysis, interpretation, and critical revision of the manuscript. Both authors read and approved the final manuscript.

Acknowledgement

The authors gratefully acknowledge the faculty and staff of the Department of Clinical Foundations and the standardized patient team for their contributions to the delivery and administration of the clinical-skills OSCEs included in this study.

Data Availability

The datasets analysed during the current study were derived from Ross University School of Medicine examination reports and are not publicly available because they contain student assessment information subject to institutional confidentiality requirements. The datasets are available from the corresponding author upon reasonable request, subject to permission from Ross University School of Medicine and applicable data protection requirements.

References

1. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J. 1975;1(5955):447–51. https://doi.org/10.1136/bmj.1.5955.447.
2. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The objective structured clinical examination (OSCE): AMEE Guide 81. Part I: an historical and theoretical perspective. Med Teach. 2013;35(9):e1437–46. https://doi.org/10.3109/0142159X.2013.818634.
3. Schuwirth LWT, van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. Med Teach. 2011;33(6):478–85. https://doi.org/10.3109/0142159X.2011.565828.

4. van der Vleuten CPM, Schuwirth LWT, Driessen EW, Dijkstra J, Tigelaar D, Baartman LKJ, et al. A model for programmatic assessment fit for purpose. Med Teach. 2012;34(3):205–14. https://doi.org/10.3109/0142159X.2012.652239.
5. van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ Theory Pract. 1996;1(1):41–67. https://doi.org/10.1007/BF00596229.
6. Wise SL, DeMars CE. Low examinee effort in low-stakes assessment: problems and potential solutions. Educ Assess. 2005;10(1):1–17. https://doi.org/10.1207/s15326977ea1001_1.
7. Finn B. Measuring motivation in low-stakes assessments. ETS Res Rep Ser. 2015;2015(2):1–17. https://doi.org/10.1002/ets2.12067.
8. Cole JS, Bergin DA, Whittaker TA. Predicting student achievement for low stakes tests with effort and task value. Contemp Educ Psychol. 2008;33(4):609–24. https://doi.org/10.1016/j.cedpsych.2007.10.002.
9. Cilliers FJ, Schuwirth LWT, Adendorff HJ, Herman N, van der Vleuten CPM. The mechanism of impact of summative assessment on medical students' learning. Adv Health Sci Educ Theory Pract. 2010;15(5):695–715. https://doi.org/10.1007/s10459-010-9232-9.
10. Chisnall B, Vince T, Hall S, Tribe R. Evaluation of outcomes of a formative objective structured clinical examination for second-year UK medical students. Int J Med Educ. 2015;6:76–83. https://doi.org/10.5116/ijme.5572.a534.
11. Alkhateeb NE, Al-Dabbagh A, Ibrahim M, Al-Tawil NG. Effect of a formative objective structured clinical examination on the clinical performance of undergraduate medical students in a summative examination: a randomized controlled trial. Indian Pediatr. 2019;56(9):745–8. https://doi.org/10.1007/s13312-019-1641-0.
12. Hawkins SC, Osborne A, Schofield SJ, Pournaras DJ, Chester JF. Improving the accuracy of self-assessment of practical clinical skills using video feedback: the importance of including benchmarks. Med Teach. 2012;34(4):279–84. https://doi.org/10.3109/0142159X.2012.658897.
13. Kam BS, Yune SJ, Lee SY, Im SJ, Baek SY. Impact of video feedback system on medical students' perception of their clinical performance assessment. BMC Med Educ. 2019;19(1):252. https://doi.org/10.1186/s12909-019-1688-6.
14. Mitchell O, Cotton N, Leedham-Green K, Elias S, Bartholomew B. Video-assisted reflection: improving OSCE feedback. Clin Teach. 2021;18(4):409–16. https://doi.org/10.1111/tct.13354.
15. Bernard AW, Ceccolini G, Feinn R, Rockfeld J, Rosenberg I, Thomas L, et al. Medical students' review of formative OSCE scores, checklists, and videos improves with student–faculty debriefing meetings. Med Educ Online. 2017;22(1):1324718. https://doi.org/10.1080/10872981.2017.1324718.
16. Thelk AD, Sundre DL, Horst SJ, Finney SJ. Motivation matters: using the Student Opinion Scale to make valid inferences about student performance. J Gen Educ. 2009;58(3):129–51.

Figures

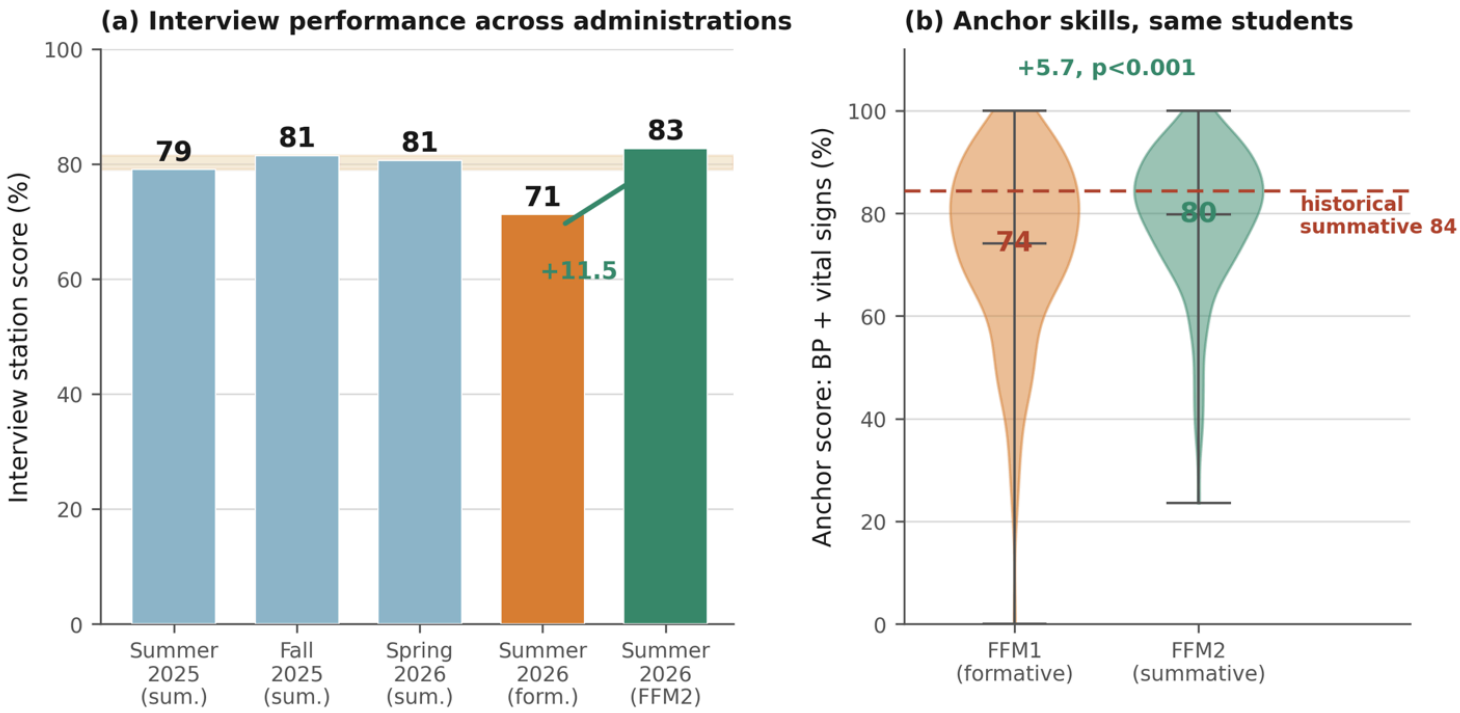


Figure 1

Interview-station performance across administrations and within-student change on the identically repeated anchor. (a) Mean interview-station scores for the three prior summative FFM1 cohorts, the formative FFM1 sitting (Summer 2026), and the same cohort's graded FFM2 sitting; the shaded band marks the historical summative range. (b) Distribution of blood-pressure and vital-signs anchor scores for the index cohort at formative FFM1 and graded FFM2 (same students); the dashed line marks the historical summative mean.

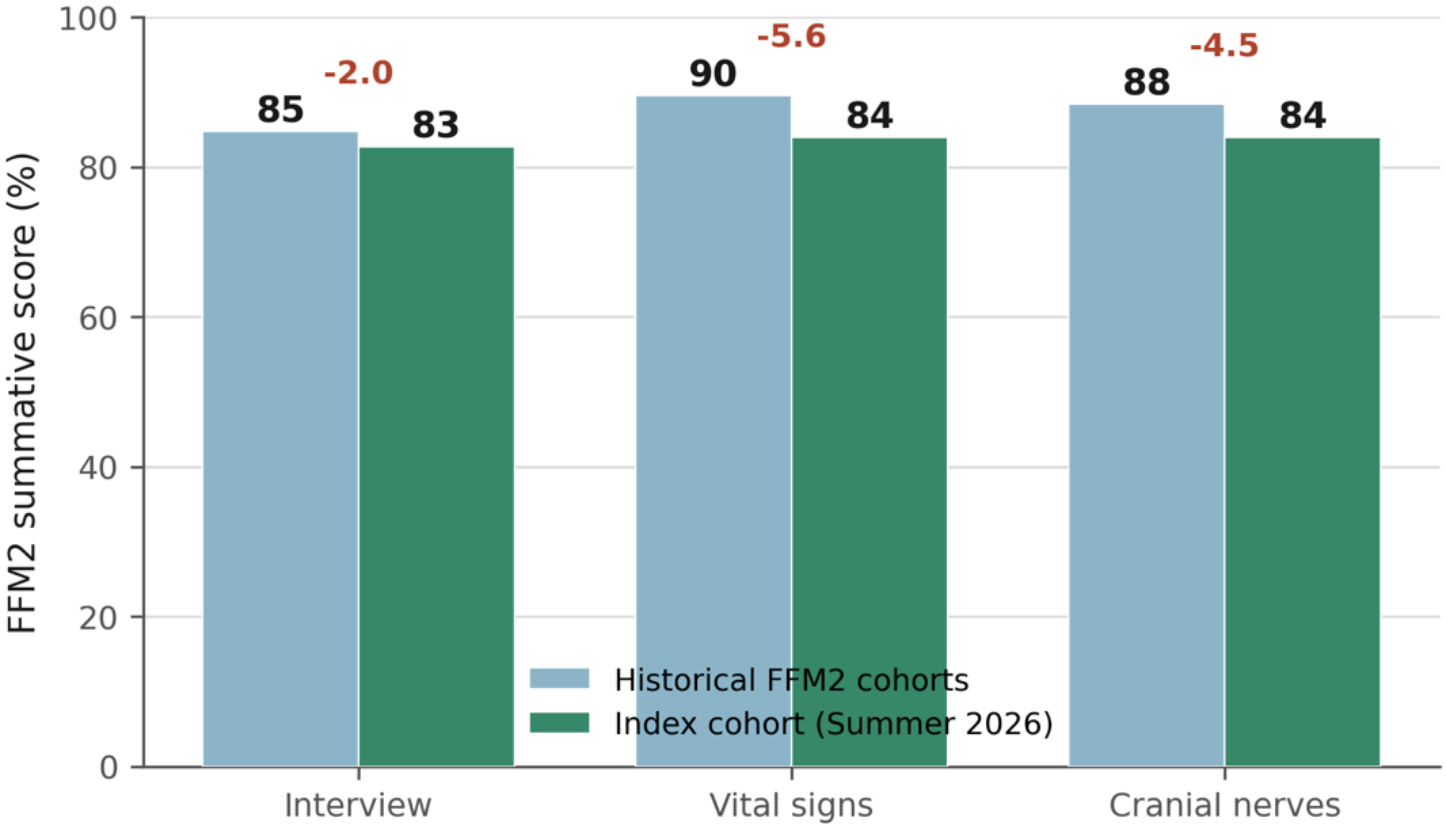


Figure 2

Index-cohort FFM2 performance versus historical FFM2 cohorts. Mean scores on the three shared, identically assessed skills (interview, vital-signs interpretation, and cranial-nerve examination), with index-cohort minus historical differences shown above each pair, illustrating near-complete but incomplete recovery.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- RUSMIRBexempt5IRB202605amendment.pdf