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.

# Instructional Revision, Facilitator-Supported Practice, and Cardiac Auscultation Performance: A Seven-Cohort Observational Study

**Charla Corbin**
Ross University School of Medicine
**Tetiana Hliebova**
Ross University School of Medicine
**Ektor Polykandriotis**
Ross University School of Medicine
**Snehal Mehta**
Ross University School of Medicine
**Liris Benjamin**
Ross University School of Medicine
**Rakesh Calton**
rcalton@rossu.edu

Ross University School of Medicine

**Research Article**

**Keywords:** cardiac auscultation, deliberate practice, supervision, self-regulated learning, objective structured clinical examination, program evaluation

**Posted Date:** November 10th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11667478/v1

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

**Additional Declarations:** Competing interest reported. Rakesh Calton is a reviewer for Medical Science Educator. Other authors do not have any competing interests.

# Instructional Revision, Facilitator-Supported Practice, and Cardiac Auscultation Performance: A Seven-Cohort Observational Study

Charla Corbin, Assistant Professor, Co-Lead Cardiology Module, Department of Clinical Foundations, Ross University School of Medicine. ORCID 0009-0003-4588-8891
Tetiana Hliebova, Associate Professor, Department of Clinical Foundations, Ross University School of Medicine. ORCID 0000-0002-0102-0914
Ektor Polykandriotis, Instructor, Department of Clinical Foundations, Ross University School of Medicine. ORCID 0009-0005-9655-5450
Snehal Mehta, Associate Professor, Department of Clinical Foundations, Ross University School of Medicine. ORCID 0000-0001-6383-3235
Liris Benjamin, Assistant Chair, Department of Clinical Foundations, Ross University School of Medicine. ORCID 0009-0006-4286-1855
Rakesh Calton, Professor and Interim Chair, Department of Clinical Foundations, and Director of Medical Simulation, Ross University School of Medicine. ORCID 0000-0002-5029-6554

Corresponding author: Rakesh Calton, rcalton@rossu.edu
Two Mile Hill, Bridgetown, Barbados

**Abstract**

Review of objective structured clinical examination (OSCE) results identified aortic sclerosis as the lowest-performing Harvey simulator station, with faculty reporting that students confused it with aortic stenosis because both present with systolic murmurs. Instruction was revised to emphasize carotid palpation and murmur radiation, the distinguishing finding, but performance did not improve. Facilitator supervision was later added to existing extended-hours practice without increasing simulator availability, allowing supervision to be examined independently of access. We analyzed seven consecutive OSCEs delivered under one curriculum and one machine-scored 14-item instrument (1,121 student-station observations, 2024 to 2026). The revision took effect in Fall 2024; four facilitators joined practice sessions from Summer 2025, intervening only when students encountered difficulty. The primary outcome was performance at the aortic sclerosis station, named in monitoring records before either intervention; mitral valve prolapse, whose diagnostic feature is independent of the revised maneuver, served as a comparison outcome. The revision alone produced no improvement, with performance across the three preceding cohorts flat to declining (83.3%, 82.2%, 81.1%). After facilitators were added, aortic sclerosis rose from 81.7% to 89.9%, a difference-in-differences of 7.87 points (95% CI 3.06 to 12.69, P = 0.001), and students scoring below 70 fell from 25.9% to 12.0%. Overall performance rose from 82.4% to 88.3% (5.65 points, 95% CI 0.87 to 10.43, bootstrap P = 0.016), while the comparison station changed by 0.40 points (P = 0.84). Facilitator-supported practice may provide educational benefit beyond access to simulation resources alone.

Keywords: cardiac auscultation; deliberate practice; supervision; self-regulated learning; objective structured clinical examination; program evaluation

## Introduction

Cardiac auscultation remains one of the least reliably acquired skills in physical examination. Multicenter studies show that competence is poor among students and improves little across training levels, with attending physicians performing only marginally better than those they supervise [1, 2]. The pattern has held for three decades, which suggests that the limiting factor is not exposure but the conditions under which the skill is learned.

Simulation can supply those conditions. The Harvey cardiopulmonary patient simulator, in use since the 1970s, reproduces pulses, precordial movement, and auscultatory findings with correct timing, intensity, and radiation [3, 4]. Simulation-based mastery learning improves auscultatory accuracy [5], repetition and deliberate practice have been identified as active ingredients [6], and a systematic review found that most studies report positive outcomes [7]. These findings sit within a broader literature identifying feedback, repetitive practice, and curricular integration as features of effective learning [8, 9].

What this literature does not resolve is a question program directors face directly. When practice and self-study are available outside scheduled teaching, does learning follow from the access itself or from a facilitator's presence? The costs differ sharply. Extending laboratory hours is inexpensive; staffing those hours is not. The common approach is to increase access, on the assumption that time on task drives skill acquisition. For a perceptual discrimination task, that assumption warrants scrutiny, since a student who repeatedly practices a misidentification has no means of self-correction.

The published evidence is mixed. A meta-analysis of self-regulated learning in simulation found that unsupervised interventions produced poorer immediate outcomes than supervised ones, with a pooled effect size of −0.34 [10]. Against this, a randomized non-inferiority trial in cardiac auscultation found self-regulated learning no worse than co-regulated learning on immediate and retention testing, though inconclusive for preparation for future learning, and raised the possibility that supervisors disrupt self-regulation by taking control of the learning situation [11]. The question remains open partly because access and supervision are usually confounded, since programs that add supervised sessions typically add hours as well, and partly because most curricular evaluation rests on uncontrolled pre-post comparison [12].

We report two interventions at a single institution that allow these influences to be distinguished. Routine monitoring identified a specific weakness. We revised the teaching of the maneuver that addresses it, and performance did not change. We later added facilitators to practice sessions while leaving opening hours unchanged, and performance improved. Because monitoring records named the station before either intervention, it serves as a target identified in advance rather than one selected from the results.

**Aim**

The aim of this study was to evaluate student performance at Harvey-based cardiac auscultation stations across seven consecutive OSCEs and to determine whether revised instruction on differentiating aortic sclerosis from aortic stenosis and the subsequent addition of facilitator-supported practice were each associated with improved performance.

## Materials and Methods

### Design

This was a retrospective observational study of consecutive OSCEs, using a non-equivalent dependent variable design [12]. Reporting follows the STROBE extension for simulation-based research, and the description of the intervention follows Table 3 of that guidance [13].

### Setting and participants

The study took place at the Simulation Institute of an offshore university delivering a United States curriculum MD program. Students move through five preclinical semesters in three overlapping cohorts each year. Participants were Semester 3 students in the Cardiology module. Each student examined one Harvey program per administration, and no student appears in more than one cohort.

### Curriculum context and analytic window

In September 2023 the school replaced a legacy curriculum, which taught cardiology across two separated semesters, with a single-module curriculum delivering cardiology as one continuous 7.5-week module, with didactic teaching reduced by half in favor of active learning. The first cohort reached Semester 3 in May 2024. The assessment instrument changed at the same point, from a facilitator-completed checklist to the machine-scored instrument described

below, and the legacy series was taught out in parallel, so it cannot serve as a like-for-like comparator. We therefore restricted the analysis to the seven administrations delivered under the single-module curriculum and the single instrument (1,121 observations). Both changes fall three cohorts before the intervention studied here.

**Identification of the target**

Harvey station results are charted after every administration as a routine quality function, and these records predate the present analysis. In the Semester 4 cohort (legacy curriculum) assessed from January to April 2024, aortic sclerosis was the lowest-performing of the five Harvey stations examined, at 79%, against 87% for normal, 84% for innocent murmur, 84% for mitral valve prolapse, and 83% for hypertension. It had also been the lowest-performing Harvey station in April 2023. Review of these records prompted discussion among clinical skills faculty, who reported that students frequently confused aortic sclerosis with aortic stenosis, plausibly because both present with systolic murmurs. The deficiency was therefore specific rather than general. Aortic sclerosis must be distinguished from aortic stenosis, and the discriminating finding is whether the systolic murmur radiates to the carotid artery. In aortic stenosis it does; in aortic sclerosis it does not [14, 15]. Faculty observed that carotid palpation during auscultation was performed inconsistently and that students rarely commented on radiation of the murmur to the neck. This observation, rather than any prior research plan, prompted the instructional revision.

**The educational activity**

Cardiac auscultation is taught as a sequence: two 50-minute physiology lectures and a two-hour workshop covering the cardiac cycle, pressure-volume relationships, and murmur genesis; a 50-minute auscultation didactic; and a 50-minute hands-on formative laboratory on nine Harvey simulators at a facilitator-to-student ratio of 1:8. Students then have access to extended-hours individual self-study practice. Performance is assessed at a summative Harvey OSCE.

The laboratory uses a fixed inspection, palpation, and auscultation protocol held constant across all nine programs, with simultaneous carotid palpation required throughout auscultation. Programs are sequenced as paired contrasts: aortic sclerosis against aortic stenosis, separated by radiation of the murmur to the carotid artery, and mitral regurgitation

against aortic regurgitation, separated by timing and shape. Holding the procedure constant is intended to reduce extraneous cognitive load so attention goes to discrimination [16], and the paired-contrast structure follows evidence that comparison supports learning of confusable diagnostic appearances [17].

**The two changes**

From the Fall 2024 cohort, the teaching of carotid palpation was revised. Students were required to palpate the carotid artery throughout auscultation, to listen at the carotid artery for the systolic murmur, and to state explicitly whether the murmur radiated to the neck. The change was directed at the aortic sclerosis and aortic stenosis discrimination and is documented in the student study guide.

After three successive cohorts showed no improvement, dedicated supervision of extended-hours self-study sessions was introduced from the Summer 2025 cohort. Four clinical skills facilitators, all graduates of the program, were available on weekdays and during Saturday sessions. Session availability and opening hours did not change. What changed was the presence of a facilitator who could observe practice and explain findings when a student struggled or could not distinguish two similar sounds.

Facilitator support was provided on an as-needed basis rather than through scheduled instructional sessions. Facilitators did not deliver structured teaching but intervened selectively when students encountered difficulty during independent practice, calibrating support to observed difficulty. This corresponds to the contingency principle in scaffolding theory [18, 19].

**Outcome measure and instrument**

The outcome was the percentage score at the Harvey station OSCE. Students receive 30 seconds of preparation time followed by 10 minutes to perform the examination and record responses on a tablet in CAE LearningSpace. No examiner is present, and scoring is automated. The instrument has 14 items totaling 100 points: 11 covering inspection, palpation, and auscultatory findings by area, including carotid radiation (75 points); one requiring the best diagnosis (15 points); and two on confirmatory investigation and associated findings (10 points). Because every item is worth 5, 10, or 15 points, all attainable totals are multiples of 5. In the analyzed data 1,110 of 1,121 scores (99.0%) conform to this key; the remaining 11 are

treated as artifacts in a sensitivity analysis. The instrument was unchanged across all seven cohorts.

Because scoring is automated with no examiner present, examiner variance is eliminated by design rather than estimated, and examiner drift cannot account for change across the series.

**Comparison outcome**

Mitral valve prolapse (MVP) was selected as a negative control outcome [20] because its principal diagnostic feature, a midsystolic click, is independent of the carotid palpation and murmur radiation maneuver targeted by the instructional revision. The station was taught within the same educational environment, administered during the same assessment sessions, and evaluated using the same automated scoring instrument. Although the instructional revision was not expected to influence performance at this station directly, facilitator-supported practice could plausibly improve recognition of the midsystolic click through corrective feedback. Consequently, MVP represents a comparison outcome with limited relevance to the targeted instructional content rather than a definitively unaffected control for the supervision intervention. The rationale for distinguishing this station was supported by instructional materials predating the intervention; however, its selection as a comparison outcome occurred during the analysis rather than prospectively.

The station nonetheless excludes influences acting on all stations together, such as cohort ability, familiarity with test items, or a change in scoring, since any of these would have raised performance at this station as well.

**Statistical analysis**

Analyses used Python 3 with statsmodels. The primary outcome was performance at the aortic sclerosis station. The estimator was a difference in differences: an ordinary least squares model of station percentage on station fixed effects, a post-supervision indicator, and the interaction of that indicator with an indicator for the targeted station. A secondary analysis pooled the eight non-comparison stations. Standard errors were clustered by cohort. Seven clusters are too few for reliable asymptotic inference, which can lead cluster-robust tests to over-reject, so we also report wild cluster bootstrap P values using Rademacher weights under the imposed null with 3,999 replications and a fixed random seed [21].

A causal interpretation of the difference-in-differences estimate would require that, absent the intervention, performance at the targeted and comparison stations would have followed parallel trajectories. This assumption cannot be tested here, because aortic sclerosis was examined in only one pre-supervision cohort. Differences between stations in diagnostic complexity, baseline performance, and responsiveness to feedback may also influence how each changes over time. The estimates are therefore interpreted as adjusted observational comparisons rather than causal effects.

Sensitivity analyses were station-level estimates with Holm correction, restriction to the four stations given in all seven cohorts, exclusion of the 11 off-key scores, rank transformation, and a derived diagnosis measure. The derived measure exploits the scoring key: the highest total attainable with the 15-point diagnosis item wrong is 85, so any score of 90 or above establishes that the diagnosis was correct. One student recorded scores at two stations, a protocol deviation under a design in which each student sits exactly one station; that student's records were excluded. Station-cohort cells containing fewer than five observations are suppressed from station-level tables but retained in pooled models. The deidentified dataset and the summary tables reported here are provided as Supplementary Data.

**Use of artificial intelligence**

A large language model (Claude, Anthropic) assisted with the extraction of data from spreadsheet files, with statistical code, was used. All analyses were specified by the authors, all outputs were checked against the source data, and the authors take full responsibility for the content.

**Ethics**

The study was exempted from IRB review vide approval letter RUSMIRB exempt #6-IRB2026_06 The analysis used deidentified results generated as routine program evaluation. No individual student is identifiable.

**Results**

**The instructional revision alone**

The Fall 2024 cohort was the first to receive the revised teaching. At the aortic sclerosis station, that cohort scored a mean of 81.7% (SD = 20.2, n = 27), with 25.9% of students below 70.

Across all stations, performance in the three cohorts before supervision was 83.3%, 82.2%, and 81.1% (Table 1). The instructional revision was not followed by improvement at the station it targeted or overall.

**Primary outcome: aortic sclerosis**

After facilitators were added, performance at the aortic sclerosis station rose to 89.9% (SD = 14.7, n = 92), a raw difference of 8.3 points (Welch P = 0.056, d = 0.47). In the difference-in-differences model the station gained 7.87 points against the comparison station (95% CI 3.06 to 12.69, P = 0.001). Students scoring below 70 fell from 25.9% to 12.0%, and the proportion whose score established a correct diagnosis rose from 55.6% to 79.3%. The cohort sequence was 81.7%, then 84.6%, 88.4%, 95.6%, and 92.2% (Fig. 1). The pre-supervision estimates rest on a single administration, because the station was not examined in the Summer 2024 or Spring 2025 cohorts (Fig. 2).

**Aortic stenosis, the other half of the pair**

Aortic stenosis station performance did not change. It scored 87.2% (SD = 13.8, n = 56) before supervision and 88.4% (SD = 13.4, n = 87) after, a difference of 1.2 points (P = 0.62; difference in differences 0.78, P = 0.80). Performance was already near the top of the distribution, with 68% of students at or above 90% beforehand. Taken as a pair, the two stations the instruction addressed show a combined difference in differences of 3.47 points (95% CI −3.83 to 10.78, P = 0.35).

**All stations**

Pooled performance rose from 82.4% (n = 453, SD = 17.0) to 88.3% (n = 668, SD = 16.1), a difference of 5.9 points (95% CI 3.9 to 7.9). Students scoring below 70 fell from 19.0% to 11.2%. The eight non-comparison stations gained 5.65 points more than the comparison station (95% CI 0.87 to 10.43, P = 0.021; bootstrap P = 0.016). The coefficient estimating change at the comparison station was 0.40 points (95% CI −3.63 to 4.44, P = 0.84). The shift was concentrated in the lower part of the distribution: the two distributions are nearly identical above 90%, while the proportion below the pass threshold fell by about two fifths (Fig. 3).

**Station-level and sensitivity analyses**

Station-level estimates ranged from 10.58 points for aortic regurgitation to 0.78 for aortic stenosis (Table 2, Fig. 4). No individual station survived Holm correction, so these analyses are descriptive. Restricting to the four stations given in every cohort gave 5.45 points (95% CI 1.95 to 8.96, P = 0.002), and excluding the 11 off-key scores gave 5.05 points (95% CI 0.26 to 9.84, P = 0.039). Rank transformation did not reproduce the pooled effect (P = 0.136). The below-70 outcome moved favorably but imprecisely (OR = 0.57, 95% CI 0.22 to 1.44, P = 0.23). Across all stations, the proportion whose score established a correct diagnosis rose from 46.6% to 69.7% (OR = 1.93, 95% CI 0.90 to 4.13, P = 0.09). Table 3 summarizes these analyses.

**Discussion**

Two findings emerged. Revising how a discriminating maneuver was taught, at the station our own monitoring had identified as lowest-performing, was not followed by improvement. Introducing facilitators while leaving practice hours unchanged was followed by an 8.3-point gain at that station and a 5.9-point gain overall, sustained across four cohorts, while a comparison station within the same examination did not change.

The first finding was unexpected and is the one most worth reporting. Content can be revised carefully, delivered by faculty who understand the discrimination, and documented in the materials students receive, and performance may remain unchanged if students then study alone.

**Access is not the active ingredient**

The pre-supervision condition provides the most informative comparison. Students had the simulators, the self-study hours, the study guide, and the audio and video materials, yet three successive cohorts showed no improvement in performance. Performance changed only when a facilitator was present during those same hours. This aligns with the original specification of deliberate practice, in which feedback and coaching are constitutive rather than optional [9, 22].

Three frameworks bear on this and are worth separating. Deliberate practice specifies repetition with feedback against a defined standard [22]. Self-regulated learning concerns the learner's own planning, monitoring, and evaluation of practice [10]. Contingent scaffolding

describes support calibrated to observed difficulty and withdrawn as competence develops [18, 19]. Our intervention is closest to the third. Facilitators set no goals, delivered no structured teaching, and intervened only when a student encountered difficulty, so the learner retained control of the session throughout.

The mechanism we propose is the limit of self-assessment. A student practicing alone has no external reference for what they believe they heard, so repetition may reinforce error as readily as accuracy. The facilitator supplies the corrective signal, functioning less as an instructor than as a navigation aid: the learner still steers but is told when the route has been missed. This framing raises a question the present design cannot answer: whether the benefit derives from the facilitator's presence or from the timely corrective feedback that presence makes available. The two are confounded here, and they carry very different resource implications.

The distribution of the change supports this reading. Scores above 90% were essentially unaltered, while the proportion below the pass threshold fell by about two fifths (Fig. 3). A student already identifying the finding correctly gains nothing from confirmation. A student identifying it incorrectly, with no way of detecting the error, gains the correction itself. This is what the rank-transformed analysis reflects, and it is a prediction of the mechanism rather than a shortcoming of the estimate.

**Why only half the pair changed**

Instruction addressed two stations and only one improved. We think the asymmetry is informative. In aortic stenosis the murmur radiates to the carotid artery, so the maneuver produces a positive finding, and it is accompanied by other cues, including the harsh ejection quality, the crescendo-decrescendo shape, and an audible fourth heart sound. Students scored 87.2% there beforehand, which left limited room to improve and reduced reliance on the maneuver.

In aortic sclerosis the diagnostic signal is the absence of murmur radiation. A student cannot confirm an absence by listening more closely; they must perform the maneuver, know where to listen, and interpret a negative result correctly. Nothing in the environment distinguishes correct listening that found nothing from incorrect listening, which is exactly the judgment an unsupervised learner cannot verify. A facilitator can confirm it in seconds. If this reading is right,

facilitator support may add least where the diagnostic cue is positive and redundant and most where it is a negative finding the learner cannot independently check. We offer this as a hypothesis generated by the data rather than tested by it.

Considered as a single contrast, the two stations the instruction addressed show no significant gain (3.47 points, 95% CI −3.83 to 10.78, P = 0.35). We report the aortic sclerosis station as the primary outcome because the institutional monitoring named it before either change, not because it performed best. Had the pair been specified as the outcome, this study would have returned a null result.

Two stations that the revised instruction did not address, aortic regurgitation and mitral stenosis, gained more than the target station. This argues against a narrowly specific mechanism. The more defensible reading is that facilitator presence helped across murmur discrimination generally, with the carotid maneuver as one instance, and that the comparison station marks the boundary of that benefit rather than the limit of a single taught cue.

**Reconciling divergent evidence**

Our findings align with meta-analytic evidence that unsupervised training yields poorer immediate outcomes [10] and appear at first to contradict the non-inferiority trial [11]. Differing definitions of unsupervised learning may explain the discrepancy. In that trial the self-regulated group received structured cases, guidance, and assessment, whereas our pre-supervision condition offered only unstructured access during open hours. Self-regulated learning may substitute for supervision when it is appropriately scaffolded, while unstructured access substitutes for neither. Contingent scaffolding also differs from the supervision that trial describes, since the learner retains control until performance falters. Our facilitators were recent graduates rather than senior faculty, which may have lowered the threshold for students to express uncertainty [23], and sessions were voluntary and ungraded.

**The comparison station as a finding**

The comparison station carries identical items and point allocations, differing only in the program presented. Its defining feature, the midsystolic click, is worth 5 points, but a student who misses it would be unlikely to reach the correct diagnosis, which carries 15 points. Roughly one fifth of the available score therefore depends on a brief perceptual event that paired-

contrast teaching does not address. This leads to a testable prediction: a dedicated click-recognition segment should improve performance at this station while leaving the murmur stations unchanged, reversing the roles of target and comparison.

**Limitations**

This study is nonrandomized and retrospective, and an unmeasured change coinciding with the intervention and affecting only murmur stations would produce the same pattern. The pre-supervision estimate for the primary outcome rests on a single administration of 27 students, because aortic sclerosis was not examined in two of the three pre-supervision cohorts; the raw difference does not reach conventional significance, and the difference-in-differences estimate draws its precision from the comparison station rather than from an extended baseline. The parallel-trends assumption underlying the estimator cannot be tested for the same reason. We identified the comparison station during analysis rather than specifying it in advance, and facilitator support could plausibly have affected it, so it is not a definitively unaffected control. Practice dose was not recorded, which precludes any dose-response analysis and means the estimate reflects an intention-to-treat contrast at the cohort level. The rank-transformed analysis did not reproduce the pooled effect, indicating that the linear estimate is driven largely by change in the lower part of the distribution. Station-level analyses are underpowered, and none survives correction. Item-level responses were unavailable, so we propose the click-gating mechanism rather than demonstrate it. Finally, this study examines one skill in one program, and the outcome reflects simulator performance rather than performance with patients [24].

**Implications**

For program directors, cost the supervision rather than the access, and treat extended opening hours without facilitation as untested. For evaluators, check before running an uncontrolled pre-post comparison whether the assessment already contains a station the intervention should not affect. For simulation programs, routine station-level monitoring is what made this evaluation possible and what named the target in advance; the comparison station can be nominated from the instructional design when a change is planned and registered in the minutes of the review that approves it.

Two questions follow. The first is dose, which requires attendance data we did not have and are now collecting. The second is whether the verification function must be human. If the operative element is a timely signal that the learner has misidentified a sound, it may be deliverable by automated feedback at a fraction of the staffing cost.

A prospective three-arm comparison would separate these. With practice exposure held equal, learners would be allocated to independent simulator practice, facilitator-supported practice, or independent practice with structured automated corrective feedback. Diagnostic accuracy, recorded practice dose, and delayed retention would be the outcomes. Such a design would establish whether human supervision is required or whether reliable, timely feedback is sufficient, a distinction with direct consequences for the cost of extended laboratory hours.

**Conclusions**

Revising how a discriminating maneuver was taught was not associated with improved performance at the station our monitoring had identified as lowest-performing. Performance improved only after facilitators were present while students practiced, by 8.3 points at that station and 5.9 points overall, while a comparison station within the same examination did not change. These findings suggest that facilitator-supported practice may provide educational benefit beyond access to simulation resources alone. Prospective research is needed to establish the relative contributions of supervision, practice frequency, and feedback.

**Declarations**

**Ethics approval and consent to participate**

Institutional IRB exempted this study vide #6-IRB2026_06.

The analysis used deidentified results generated as routine program evaluation. No individual student is identifiable.

**Consent to participate and for publication**

Not applicable. The analysis used deidentified institutional assessment records.

**Data availability**

Deidentified station-level data and the summary tables reported in this article are provided as Supplementary Data. Analysis code is available from the corresponding author on reasonable

request. Full student assessment records are held under institutional data governance and are not publicly available.

**Conflict of interest**

RC designed the educational activity under evaluation, developed the assessment instrument, and conducted the analysis. Four factors mitigate this. The data were collected routinely for program quality purposes rather than for research. The instrument was unchanged across the analytic window. Scoring was automated with no examiner present. The pedagogical rationale for the comparison station is documented in teaching materials that predate the intervention, although the station was identified as a comparison outcome during analysis rather than before it.

**Funding**

None.

**Authors' contributions**

CC, as team lead, designed the educational activity. RC specified the analysis, performed the statistical analysis, and drafted the manuscript. SM contributed critical revision and edits. EP contributed to methodological interpretation and critical revision. LB and TH contributed to study framing and critical revision. All authors participated in delivering the educational activity under evaluation, contributed to revision for important intellectual content, and read and approved the final version.

**Acknowledgements**

The authors are grateful to the clinical skills facilitators who supervised extended-hours practice, to the simulation operations staff who maintained the equipment and ran the examinations, and to the cardiology module faculty for their support throughout.

## Tables

### Table 1 Mean station performance by cohort

| Cohort | Period | n | Mean (%) | SD |
|---|---|---|---|---|
| Summer 2024 | Pre-supervision | 179 | 83.3 | 16.8 |
| Fall 2024 | Pre-supervision | 159 | 82.2 | 16.8 |
| Spring 2025 | Pre-supervision | 115 | 81.1 | 17.6 |
| Summer 2025 | Post-supervision | 165 | 87.5 | 16.4 |
| Fall 2025 | Post-supervision | 158 | 88.2 | 16.6 |
| Spring 2026 | Post-supervision | 116 | 88.8 | 15.4 |
| Summer 2026 | Post-supervision | 229 | 88.7 | 15.8 |

Note: Values are the mean of all stations administered in that cohort. SD standard deviation. The instructional revision took effect in the Fall 2024 cohort; facilitator supervision began with the Summer 2025 cohort. Cohort size varies because intake alternates between larger and smaller semesters.

### Table 2 Station-level difference in differences against the comparison station

| Station | n pre | n post | Mean pre | Mean post | DiD | 95% CI | P | P Holm |
|---|---|---|---|---|---|---|---|---|
| Aortic regurgitation | 55 | 43 | 77.2 | 88.2 | 10.58 | 0.73 to 20.43 | 0.035 | 0.284 |
| Mitral stenosis | 58 | 67 | 71.4 | 80.3 | 8.49 | −0.72 to 17.70 | 0.071 | 0.494 |
| Aortic sclerosis | 27 | 92 | 81.7 | 89.9 | 7.87 | −1.81 to 17.56 | 0.111 | 0.663 |
| Mitral regurgitation | 57 | 77 | 78.6 | 85.1 | 6.08 | −1.83 to 13.99 | 0.131 | 0.663 |
| Normal | 38 | 70 | 84.2 | 90.0 | 5.39 | −3.67 to 14.44 | 0.243 | 0.882 |
| Hypertension | 37 | 77 | 87.4 | 93.2 | 5.34 | −3.23 to 13.92 | 0.221 | 0.882 |
| Innocent murmur | 67 | 78 | 89.0 | 92.2 | 2.87 | −5.19 to 10.93 | 0.484 | 0.968 |
| Aortic stenosis | 56 | 87 | 87.2 | 88.4 | 0.78 | −7.07 to 8.62 | 0.846 | 0.968 |
| Mitral valve prolapse | 58 | 79 | 85.4 | 85.8 | reference | — | — | — |

Note: Stations are ordered by the size of the estimate. DiD difference in differences, expressed in percentage points relative to the comparison station. Mitral valve prolapse served as the comparison outcome and is the reference category. P Holm gives the P value after Holm correction across the eight comparisons; no station survives correction, so these results are descriptive and hypothesis-generating. The confidence interval for aortic sclerosis differs from that reported for the primary analysis because the primary model used cohort-clustered standard errors, whereas this table reports conventional standard errors to keep the eight comparisons on a common basis.

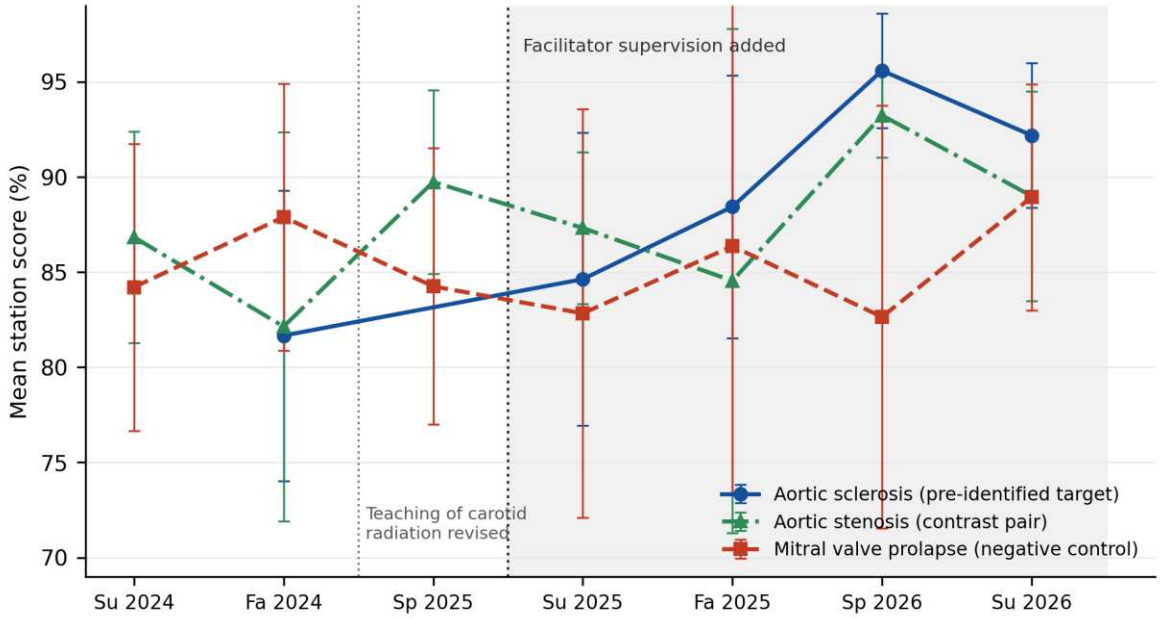
**Table 3 Primary and sensitivity analyses**

| Analysis | n | Estimate | 95% CI | P |
|---|---|---|---|---|
| Aortic sclerosis, primary outcome | 1,121 | 7.87 | 3.06 to 12.69 | 0.001 |
| All stations, cohort-clustered SE | 1,121 | 5.65 | 0.87 to 10.43 | 0.021 |
| All stations, wild cluster bootstrap | 1,121 | 5.65 | — | 0.016 |
| Comparison station change | 1,121 | 0.40 | −3.63 to 4.44 | 0.844 |
| Aortic pair combined | 399 | 3.47 | −3.83 to 10.78 | 0.352 |
| Stations given in all seven cohorts | 511 | 5.45 | 1.95 to 8.96 | 0.002 |
| Excluding 11 off-key scores | 1,110 | 5.05 | 0.26 to 9.84 | 0.039 |
| Rank-transformed outcome | 1,121 | 0.065 | — | 0.136 |
| Scoring below 70, odds ratio | 1,121 | 0.57 | 0.22 to 1.44 | 0.231 |
| Derived diagnosis correct, odds ratio | 1,121 | 1.93 | 0.90 to 4.13 | 0.091 |

Note: The first two rows are the primary and secondary analyses; the remainder are sensitivity analyses. Estimates are percentage points except where an odds ratio is indicated. SE standard error. The wild cluster bootstrap used Rademacher weights under the imposed null with 3,999 replications and a fixed random seed, appropriate here because seven clusters are too few for reliable asymptotic inference. The rank-transformed analysis does not reproduce the pooled effect, which indicates that the linear estimate is driven largely by movement in the lower part of a bounded, left-skewed distribution.

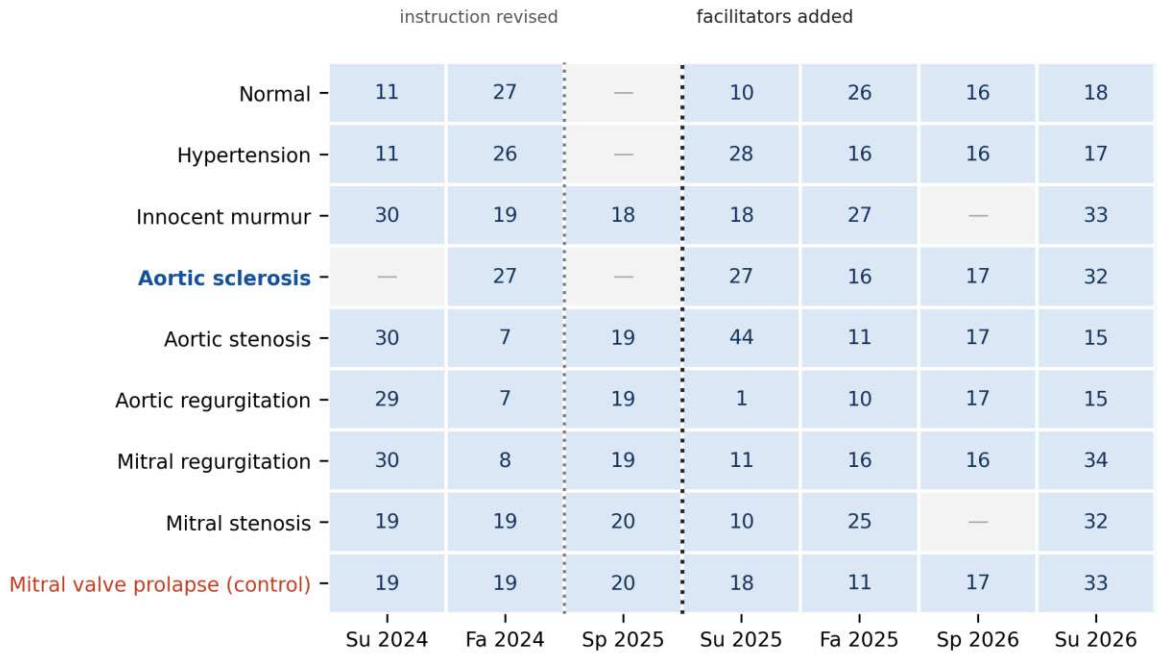
## Figures

**Fig. 1 Performance at the target station, the contrast-pair station, and the comparison outcome**



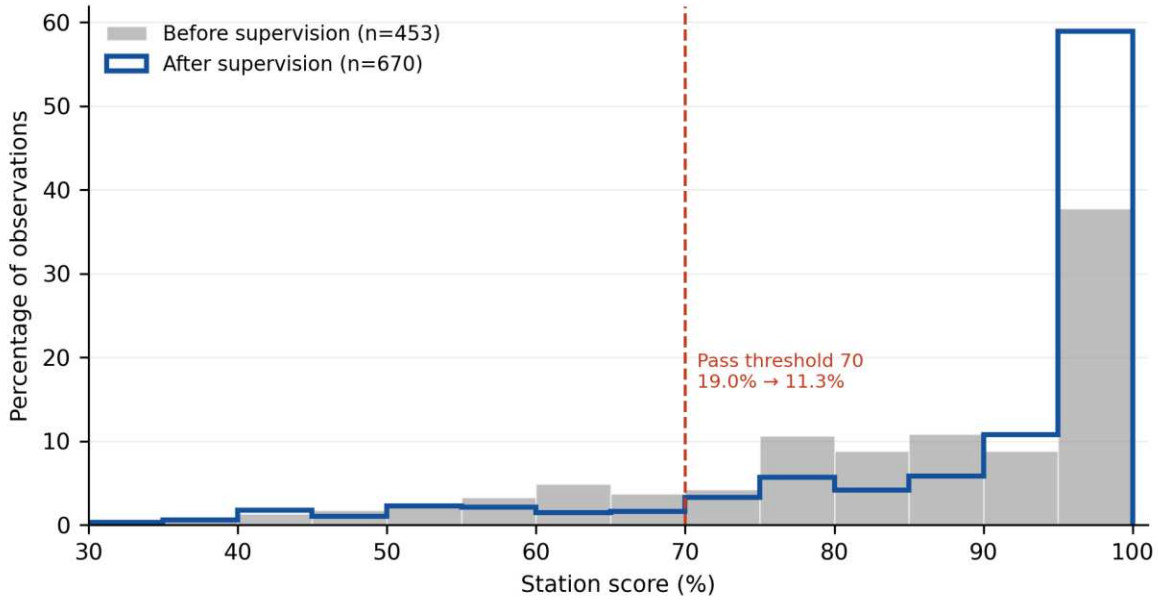

Note: Mean score across seven consecutive administrations at the station identified in advance as the target (aortic sclerosis), at the other member of the taught contrast pair (aortic stenosis), and at the comparison outcome (mitral valve prolapse). Error bars are 95% confidence intervals. The first dotted line marks revision of the teaching of carotid palpation and murmur radiation from the Fall 2024 cohort. The second marks the addition of facilitator supervision from Summer 2025, with opening hours unchanged. Aortic sclerosis was not examined in the Summer 2024 or Spring 2025 cohorts, so no point is plotted for those administrations. Cells carry 11 to 44 observations, so individual intervals are wide.

**Fig. 2 Stations administered in each cohort**



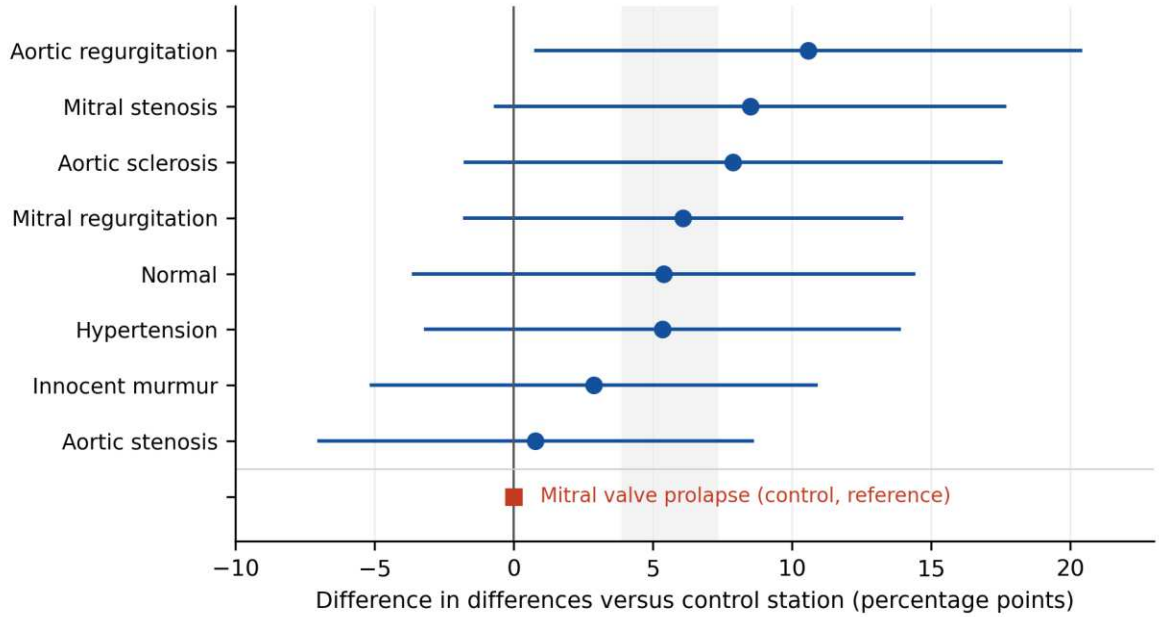

Note: Number of students examined at each station in each administration. Shaded cells with a dash indicate that the station was not administered in that cohort, which is a structural absence rather than missing data. The target station is shown in bold and the comparison outcome in red. The dotted lines mark the two interventions. The figure shows the basis of the primary analysis and makes clear that the pre-supervision estimate for aortic sclerosis rests on a single administration.

**Fig. 3 Distribution of station scores before and after facilitator supervision**




Note: Distribution of all station scores in the three cohorts before supervision and the four cohorts after. The two distributions are nearly identical in the upper range, while the proportion of scores below the pass threshold of 70 falls from 19.0% to 11.2%. This supports the interpretation of the rank-transformed sensitivity analysis: the gain is concentrated among lower-performing students rather than distributed evenly.

**Fig. 4 Station-level difference in differences against the comparison station**




Note: Point estimates with 95% confidence intervals for each station, expressed in percentage points against the mitral valve prolapse comparison. The shaded band marks the pooled estimate of 5.65 points. Stations are ordered by the size of the estimate. No individual station survives Holm correction across the eight comparisons, so these estimates are descriptive.

## References


1. Mangione S, Nieman LZ. Cardiac auscultatory skills of internal medicine and family practice trainees: a comparison of diagnostic proficiency. JAMA. 1997;278(9):717-22. doi:10.1001/jama.1997.03550090041030

2. Vukanovic-Criley JM, Criley S, Warde CM, Boker JR, Guevara-Matheus L, Churchill WH, Nelson WP, Criley JM. Competency in cardiac examination skills in medical students, trainees, physicians, and faculty: a multicenter study. Arch Intern Med. 2006;166(6):610-6. doi:10.1001/archinte.166.6.610

3. Gordon MS. Cardiology patient simulator: development of an animated manikin to teach cardiovascular disease. Am J Cardiol. 1974;34(3):350-5. doi:10.1016/0002-9149(74)90038-1

4. Ewy GA, Felner JM, Juul D, Mayer JW, Sajid AW, Waugh RA. Test of a cardiology patient simulator with students in fourth-year electives. J Med Educ. 1987;62(9):738-43. doi:10.1097/00001888-198709000-00005

5. Butter J, McGaghie WC, Cohen ER, Kaye M, Wayne DB. Simulation-based mastery learning improves cardiac auscultation skills in medical students. J Gen Intern Med. 2010;25(8):780-5. doi:10.1007/s11606-010-1309-x

6. Barrett MJ, Lacey CS, Sekara AE, Linden EA, Gracely EJ. Mastering cardiac murmurs: the power of repetition. Chest. 2004;126(2):470-5. doi:10.1378/chest.126.2.470

7. Patrizio HA, Phyu R, Kim B, Brolis NV. Utilization of simulation to teach cardiac auscultation: a systematic review. Cureus. 2023;15(7):e41567. doi:10.7759/cureus.41567

8. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach. 2005;27(1):10-28. doi:10.1080/01421590500046924

9. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. Acad Med. 2011;86(6):706-11. doi:10.1097/ACM.0b013e318217e119

10. Brydges R, Manzone J, Shanks D, Hatala R, Hamstra SJ, Zendejas B, Cook DA. Self-regulated learning in simulation-based training: a systematic review and meta-analysis. Med Educ. 2015;49(4):368-78. doi:10.1111/medu.12649

11. Lorello GR, Hodwitz K, Issenberg SB, Brydges R. Relinquishing control? Supervisor co-regulation may disrupt students' self-regulated learning during simulation-based training. Adv Health Sci Educ Theory Pract. 2024;29(1):9-25. doi:10.1007/s10459-023-10244-9

12. Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin; 2002.

13. Cheng A, Kessler D, Mackinnon R, Chang TP, Nadkarni VM, Hunt EA, Duval-Arnould J, Lin Y, Cook DA, Pusic M, Hui J, Moher D, Egger M, Auerbach M. Reporting guidelines for health care simulation research: extensions to the CONSORT and STROBE statements. Simul Healthc. 2016;11(4):238-48. doi:10.1097/SIH.0000000000000150

14. Etchells E, Bell C, Robb K. Does this patient have an abnormal systolic murmur? JAMA. 1997;277(7):564-71. PMID: 9032164.

15. Etchells E, Glenns V, Shadowitz S, Bell C, Siu S. A bedside clinical prediction rule for detecting moderate or severe aortic stenosis. J Gen Intern Med. 1998;13(10):699-704. doi:10.1046/j.1525-1497.1998.00207.x

16. van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: design principles and strategies. Med Educ. 2010;44(1):85-93. doi:10.1111/j.1365-2923.2009.03498.x

17. Kok EM, de Bruin ABH, Robben SGF, van Merriënboer JJG. Learning radiological appearances of diseases: does comparison help? Learn Instr. 2013;23:90-7. doi:10.1016/j.learninstruc.2012.07.004
18. Wood D, Bruner JS, Ross G. The role of tutoring in problem solving. J Child Psychol Psychiatry. 1976;17(2):89-100. doi:10.1111/j.1469-7610.1976.tb00381.x
19. van de Pol J, Volman M, Beishuizen J. Scaffolding in teacher-student interaction: a decade of research. Educ Psychol Rev. 2010;22(3):271-96. doi:10.1007/s10648-010-9127-6
20. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology. 2010;21(3):383-8. doi:10.1097/EDE.0b013e3181d61eeb
21. Cameron AC, Gelbach JB, Miller DL. Bootstrap-based improvements for inference with clustered errors. Rev Econ Stat. 2008;90(3):414-27. doi:10.1162/rest.90.3.414
22. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med. 2004;79(10 Suppl):S70-81. doi:10.1097/00001888-200410001-00022
23. Hadwin AF, Järvelä S, Miller M. Self-regulation, co-regulation, and shared regulation in collaborative learning environments. In: Schunk DH, Greene JA, editors. Handbook of self-regulation of learning and performance. 2nd ed. New York: Routledge; 2017. p. 83-106. doi:10.4324/9781315697048-6
24. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65(9 Suppl):S63-7. doi:10.1097/00001888-199009000-00045

## Supplementary Files

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

- SupplementaryData.xlsx