Additional file 1

Per-rater OSCE comparisons, inter-rater reliability (ICC forest plot), and misclassification sensitivity analysis for questionnaire group assignment.

S1. Per-rater OSCE comparisons

Table S1 reports each rater's between-group comparison for the final total score, checklist total, four checklist domains, and the two global rating items. Tests are two-sided Mann-Whitney U; effect size is Cliff's δ with 95% bootstrap confidence intervals (1000 resamples). FDR-adjusted p values use the Benjamini-Hochberg correction within this family of 32 tests. All 12 comparisons of final total, checklist total, and D2 across the four raters favoured the experimental group.

Table S1. Per-rater OSCE comparisons (n=122 students)

Rater	Outcome	Experimental mean±SD	Control mean±SD	U	Cliff δ (95% CI)	Raw p	p_FDR
Rater A	Final total (110)	70.74±6.69	65.62±7.62	2692	+0.471 (+0.29, +0.63)	<0.001	<0.001
Rater A	Checklist total (100)	64.15±6.12	59.25±7.20	2690	+0.470 (+0.28, +0.64)	<0.001	<0.001
Rater A	D1 History & rapport (25)	18.79±2.34	17.57±2.78	2418	+0.300 (+0.09, +0.48)	0.004	0.008
Rater A	D2 Behaviour management (30)	21.72±2.30	19.77±2.54	2676	+0.462 (+0.29, +0.63)	<0.001	<0.001
Rater A	D3 Diagnostic reasoning (20)	10.00±2.36	8.83±2.57	2358	+0.267 (+0.06, +0.46)	0.010	0.019
Rater A	D4 Treatment communication (25)	13.56±2.49	13.08±2.19	2138	+0.150 (-0.04, +0.35)	0.142	0.197
Rater A	GRS-A Clinical logic (5)	3.21±0.87	3.15±0.95	1972	+0.060 (-0.13, +0.25)	0.546	0.602
Rater A	GRS-B Communication & empathy (5)	3.39±0.95	3.22±0.80	2028	+0.090 (-0.11, +0.29)	0.363	0.449
Rater B	Final total (110)	75.29±7.06	70.16±8.91	2558	+0.422 (+0.23, +0.58)	<0.001	<0.001
Rater B	Checklist total (100)	67.53±6.44	62.58±8.33	2620	+0.433 (+0.25, +0.60)	<0.001	<0.001
Rater B	D1 History & rapport (25)	19.11±2.44	18.32±2.39	2309	+0.262 (+0.06, +0.45)	0.012	0.021

Rater B	D2 Behaviour management (30)	22.89±2.22	20.60±2.71	2880	+0.548 (+0.36, +0.70)	<0.001	<0.001
Rater B	D3 Diagnostic reasoning (20)	11.15±2.09	10.38±2.75	2204	+0.185 (-0.02, +0.39)	0.074	0.108
Rater B	D4 Treatment communication (25)	14.39±2.00	13.35±2.37	2399	+0.290 (+0.10, +0.47)	0.004	0.009
Rater B	GRS-A Clinical logic (5)	3.74±0.83	3.66±0.78	1922	+0.051 (-0.13, +0.23)	0.606	0.625
Rater B	GRS-B Communication & empathy (5)	4.02±0.74	3.92±0.83	1976	+0.062 (-0.12, +0.25)	0.520	0.595
Rater C	Final total (110)	72.90±6.71	68.75±7.55	2615	+0.406 (+0.22, +0.59)	<0.001	<0.001
Rater C (gold standard)	Checklist total (100)	65.94±6.16	62.12±7.10	2595	+0.395 (+0.20, +0.58)	<0.001	<0.001
Rater C	D1 History & rapport (25)	19.03±2.29	17.83±2.94	2444	+0.314 (+0.12, +0.50)	0.002	0.006
Rater C	D2 Behaviour management (30)	22.16±2.36	20.37±2.61	2752	+0.480 (+0.29, +0.64)	<0.001	<0.001
Rater C	D3 Diagnostic reasoning (20)	11.03±1.81	10.23±2.29	2336	+0.256 (+0.05, +0.46)	0.013	0.022
Rater C (gold standard)	D4 Treatment communication (25)	13.71±1.84	13.68±2.13	1968	+0.058 (-0.14, +0.25)	0.564	0.602
Rater C	GRS-A Clinical logic (5)	3.42±0.93	3.30±0.85	1986	+0.068 (-0.13, +0.26)	0.498	0.590
Rater C	GRS-B Communication & empathy (5)	3.55±0.84	3.33±0.90	2089	+0.123 (-0.07, +0.31)	0.214	0.285
Rater D	Final total (110)	72.44±6.98	66.66±9.28	2642	+0.445 (+0.25, +0.62)	<0.001	<0.001
Rater D	Checklist total (100)	65.47±6.43	59.90±8.80	2628	+0.437 (+0.24, +0.61)	<0.001	<0.001
Rater D	D1 History & rapport (25)	19.00±2.40	18.13±2.60	2307	+0.240 (+0.03, +0.43)	0.021	0.033
Rater D	D2 Behaviour management (30)	22.06±2.52	19.45±2.98	2822	+0.517 (+0.34, +0.68)	<0.001	<0.001
Rater D	D3 Diagnostic reasoning (20)	10.94±1.95	9.53±2.91	2454	+0.341 (+0.13, +0.51)	0.001	0.003
Rater D	D4 Treatment	13.47±2.43	12.67±2.61	2232	+0.200	0.050	0.076

	communication (25)				(+0.00, +0.41)		
Rater D	GRS-A Clinical logic (5)	3.53±0.99	3.40±0.85	2028	+0.091 (-0.12, +0.29)	0.364	0.449
Rater D	GRS-B Communication & empathy (5)	3.44±0.90	3.37±0.92	1944	+0.045 (-0.14, +0.24)	0.652	0.652

S2. Inter-rater reliability (ICC(2,k))

Domain-level inter-rater agreement among the four faculty raters is summarised in Table S2 and visualised in Fig. S1. Final total ICC(2,k)=0.95 is consistent with the value reported in the main text. Domain ICCs range from 0.74 (GRS-B Communication & empathy) to 0.94 (Checklist total), with global rating items showing slightly lower agreement than checklist domains.

Table S2. ICC(2,k) by OSCE outcome

Outcome	ICC(2,k)	95% CI
Final total score (110)	0.95	0.93, 0.96
Checklist total (100)	0.94	0.92, 0.96
D1 History & rapport (25)	0.88	0.85, 0.91
D2 Behaviour management (30)	0.86	0.82, 0.89
D3 Diagnostic reasoning (20)	0.83	0.79, 0.87
D4 Treatment communication (25)	0.85	0.81, 0.88
GRS-A Clinical logic (5)	0.76	0.69, 0.82
GRS-B Communication & empathy (5)	0.74	0.67, 0.80

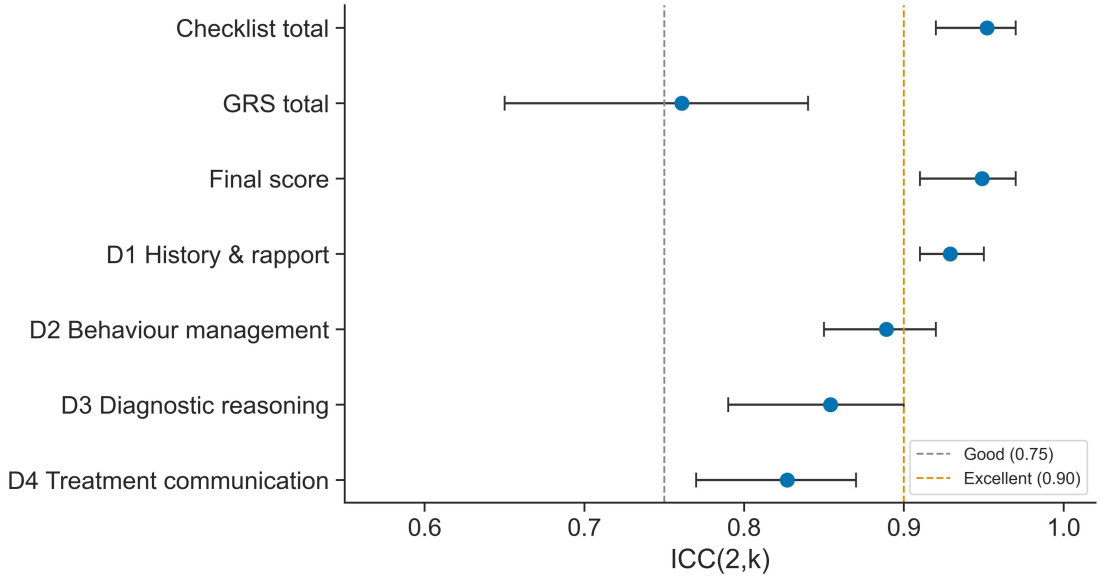

Fig. S1. Inter-rater agreement ICC(2,k) with 95% confidence intervals for OSCE outcomes.

S3. Misclassification sensitivity analysis (roster–survey n-discrepancy)

Posttest questionnaire group assignment was self-reported (experimental n=66, control n=56), whereas the OSCE used the official class roster (experimental n=62, control n=60). To examine the robustness of posttest self-efficacy group comparisons to potential misclassification, we randomly reassigned 4 experimental respondents to control and 4 control respondents to experimental in 1000 independent draws, recomputing Cliff's δ and Mann-Whitney p for each of the eight self-efficacy items in every draw. Table S3 summarises the median delta, the 2.5-97.5 percentile range, and the proportion of draws in which the raw p was below 0.05 for each item. Treatment planning remained the most robust between-group difference; the direction of the effect was preserved across all draws for every item, supporting the conclusion that misclassification of up to 4 respondents in each direction does not reverse the principal between-group findings.

Table S3. Posttest self-efficacy between-group misclassification sensitivity (±4 reassignments; 1000 draws)

Self-efficacy item	Median Cliff δ	2.5-97.5 percentile	Proportion of draws with raw p<0.05
Rapport	+0.050	(-0.049, +0.151)	0.0%
History	+0.039	(-0.056, +0.145)	0.1%
Planning	+0.256	(+0.173, +0.361)	90.0%
Risk communication	+0.108	(+0.012, +0.211)	3.8%
Emergency	+0.113	(+0.015, +0.220)	4.7%
Diagnosis	+0.135	(+0.037, +0.239)	10.7%
Pain control	+0.148	(+0.054, +0.242)	14.6%
Prevention	+0.101	(+0.006, +0.199)	2.5%