APPENDIX

Appendix A: Supplementary statistical tables

Table A1

Component-level performance by AI interaction mode

Component	No AI	AI Tutor	Direct AI
Code-related components			
Correctness	2.83 (1.46)	3.10 (1.16)	3.57 (0.73)
OOP principles	2.07 (1.17)	2.60 (0.72)	3.00 (0.00)
Code quality	0.97 (0.61)	1.37 (0.72)	1.63 (0.56)
Validation / error handling	0.63 (0.49)	0.70 (0.47)	1.00 (0.00)
Explanation quality	1.83 (1.15)	2.03 (0.96)	2.23 (1.01)
Understanding components			
Conceptual understanding	0.90 (0.31)	0.90 (0.31)	0.97 (0.18)
Applied understanding	0.90 (0.31)	0.97 (0.18)	0.93 (0.25)
Reasoning	0.93 (0.25)	0.97 (0.18)	0.97 (0.18)
Transfer	0.90 (0.31)	0.73 (0.45)	0.83 (0.38)

Note. Values are mean (SD). Code score includes correctness, OOP principles, code quality, and validation/error handling. Explanation quality is reported separately.

Table A2

Full student perceptions and reflection themes

Section	Item	N (%)
Perceived most effective mode	No AI	1 (3.3%)
	AI Tutor	21 (70.0%)
	Direct AI	8 (26.7%)
Perceived easiest mode	No AI	1 (3.3%)
	AI Tutor	9 (30.0%)
	Direct AI	20 (66.7%)
Perceived fastest mode	No AI	1 (3.3%)
	AI Tutor	2 (6.7%)
	Direct AI	27 (90.0%)
Reflection themes	Speed / time saving	10 (33.3%)
	Ease / convenience	6 (20.0%)
	Difficulty	6 (20.0%)
	Understanding	20 (66.7%)
	Learning	19 (63.3%)
	AI as help	14 (46.7%)
	AI dependence / overreliance	10 (33.3%)
	Value of No AI	7 (23.3%)
	Value of Direct AI	6 (20.0%)
	Value of AI Tutor	9 (30.0%)
	Academic integrity / ethics	4 (13.3%)
Main reflection theme	Unclear / general	7 (23.3%)
	Mixed	6 (20.0%)
	Learning	6 (20.0%)
	AI Tutor value	4 (13.3%)
	AI dependence	3 (10.0%)
	Direct AI value	2 (6.7%)
	Ease	1 (3.3%)
	Understanding	1 (3.3%)

© Yurii Kleban, 2026.

Table A3

Alignment between objective performance and student perceptions

Alignment criterion	Subjective field	Objective metric	Matching rule	Matches
Fastest mode	fastest_mode	Minimum completion time	Subjective mode is among objectively fastest mode(s)	24 (80.0%)
Most effective mode	most_effective_mode	Maximum total score	Subjective mode is among objectively highest-scoring mode(s)	18 (60.0%)
Most effective mode	most_effective_mode	Maximum code score	Subjective mode is among objectively highest code-score mode(s)	19 (63.3%)
Easiest mode	easiest_mode	Minimum perceived difficulty	Subjective mode is among objectively easiest mode(s)	20 (66.7%)

Note. Ties were handled inclusively: if more than one mode shared the best objective value for a student, a subjective response was counted as aligned if it matched any tied mode.

Table A4

Friedman tests for differences across task–mode combinations

Outcome	N	$\chi^2(2)$	p-value	Kendall's W
Time required, minutes	30	24.678	<0.001	0.411
Code score	30	11.768	0.003	0.196
Understanding score	30	1.244	0.537	0.021
Total score	30	11.347	0.003	0.189
Perceived task difficulty	30	10.634	0.005	0.177
Product–understanding gap	30	12.587	0.002	0.21
Code score per minute	30	34.101	<0.001	0.568
Understanding score per minute	30	25.966	<0.001	0.433
Total score per minute	30	29.462	<0.001	0.491

Note. All time-based analyses included 30 participants. These tests are exploratory because task and mode were paired.

Table A5

Pairwise Wilcoxon signed-rank comparisons with Holm adjustment

Outcome	Comparison	N	Mdn Δ	Mean Δ	W	Holm p	|r_rb|	Effect
Time, min	AI Tutor − No AI	30	1.5	−1.97	188	0.733	0.074	small
Time, min	**Direct AI − No AI**	30	−13.0	−14.20	22.5	**<0.001**	**0.903**	**large**
Time, min	**Direct AI − AI Tutor**	30	−13.5	−12.23	59.5	**<0.001**	**0.744**	**large**
Code score	AI Tutor − No AI	30	1	1.27	91	0.052	0.44	medium
Code score	**Direct AI − No AI**	30	1	2.7	42.5	**0.006**	**0.717**	**large**
Code score	Direct AI − AI Tutor	30	0	1.43	34.5	0.051	0.596	large
Understanding	AI Tutor − No AI	30	0	−0.07	27.5	1	0.167	small
Understanding	Direct AI − No AI	30	0	0.07	21	1	0.067	small
Understanding	Direct AI − AI Tutor	30	0	0.13	28	1	0.152	small
Total score	AI Tutor − No AI	30	1	1.2	85.5	0.089	0.43	medium
Total score	**Direct AI − No AI**	30	1	2.77	48	**0.01**	**0.68**	**large**
Total score	Direct AI − AI Tutor	30	0	1.57	65	0.089	0.486	medium
Task difficulty	AI Tutor − No AI	30	0	0	105	1	0	small
Task difficulty	**Direct AI − No AI**	30	0	−0.83	26	**0.018**	**0.696**	**large**
Task difficulty	**Direct AI − AI Tutor**	30	−1.0	−0.83	25.5	**0.007**	**0.757**	**large**
P–U gap	AI Tutor − No AI	30	1	1.33	87.5	0.072	0.417	medium
P–U gap	**Direct AI − No AI**	30	1	2.63	25	**0.005**	**0.784**	**large**
P–U gap	**Direct AI − AI Tutor**	30	0	1.3	43.5	**0.042**	**0.586**	**large**

Note. Bold = significant at α = .05 after Holm correction. Mdn Δ and Mean Δ = median and mean of within-subject differences; negative values indicate that the first-named mode had lower values. W = Wilcoxon signed-rank test statistic. |r_rb| = magnitude of the rank-biserial correlation (effect size); |r_rb| < 0.30 = small, 0.30–0.50 = medium, > 0.50 = large [34, p. 157], [29]. P–U gap = product–understanding gap. N = 30 for all comparisons. N denotes the number of complete paired observations; zero differences were omitted by the Wilcoxon procedure where applicable.

Table A6

Associations between perceived difficulty, time, and performance outcomes

Relationship	N	Spearman ρ	p-value	Interpretation
Perceived difficulty vs time required	90	0.216	0.041	Weak positive association
Perceived difficulty vs code score	90	−0.274	0.009	Higher difficulty is associated with lower code quality
Perceived difficulty vs understanding score	90	−0.175	0.099	Weak negative association
Perceived difficulty vs total score	90	−0.314	0.003	Higher difficulty is associated with lower overall performance
Perceived difficulty vs product–understanding gap	90	−0.234	0.026	Higher difficulty is associated with smaller product–understanding gap

Note. Spearman correlations were calculated using task-level observations. Time was measured in minutes.