

INTRODUCTION

- Despite the growing emphasis on competency-based education in general surgery training, existing evaluation methods remain largely subjective and susceptible to bias.
- Video-based assessment has the potential to provide more objective measures of surgical performance.

Aim:

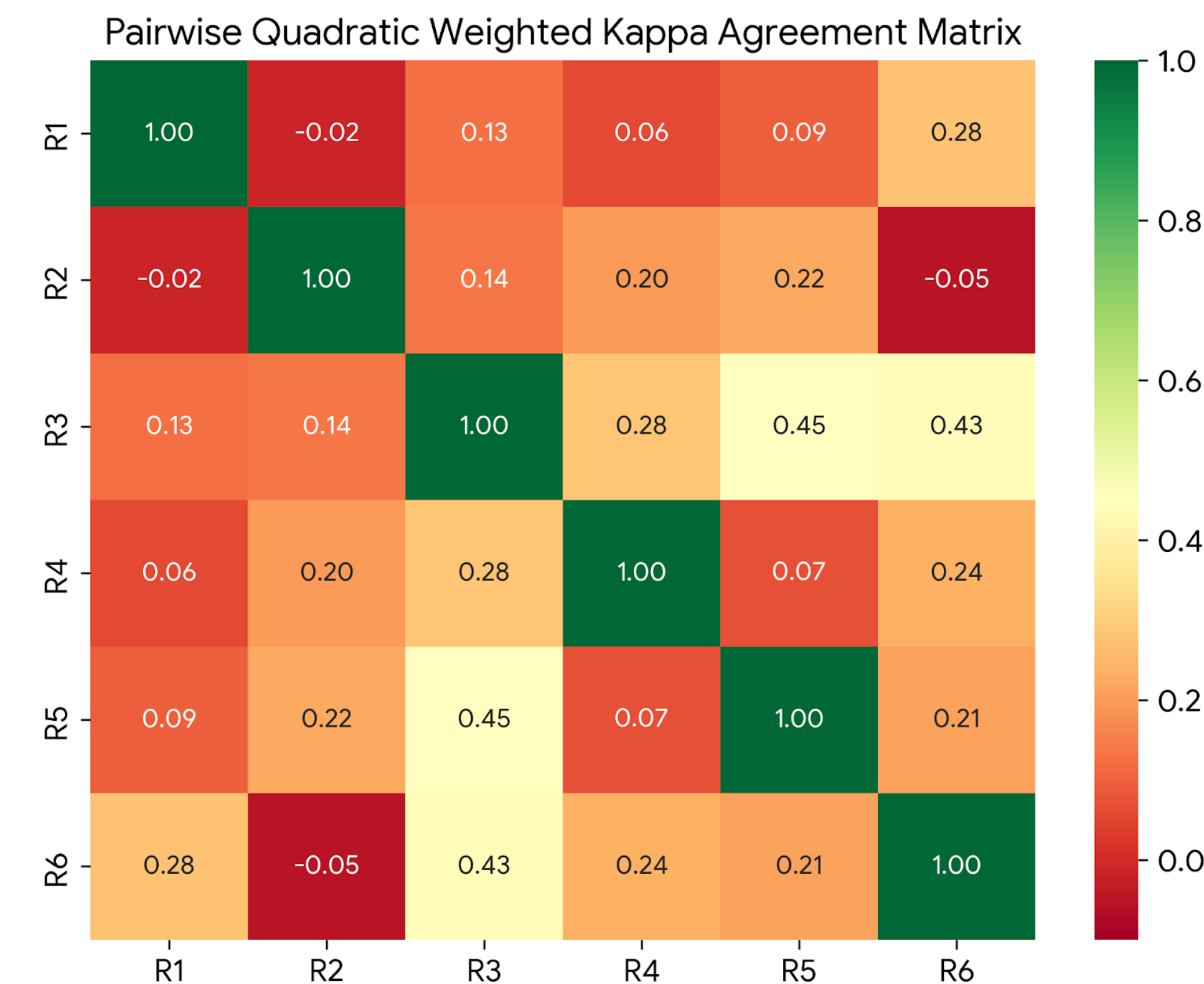
- Correlate objective assessment tool scores with both perceived and actual postgraduate year (PGY) level during video-based evaluation of laparoscopic appendectomy.

METHODS AND PROCEDURES

- Video footage was collected from laparoscopic appendectomies completed by residents under faculty supervision at a large tertiary care, academic medical center.
- The videos were edited into three 30-second clips of key portions of the procedure including port placement, dissection, and stapler division of the appendix.
- A multi-specialty group of faculty independently reviewed video footage and evaluated resident performance by using a modified Global Operative Assessment of Laparoscopic Skills (GOALS) tool and subsequently predicted the resident's level of training based on performance.

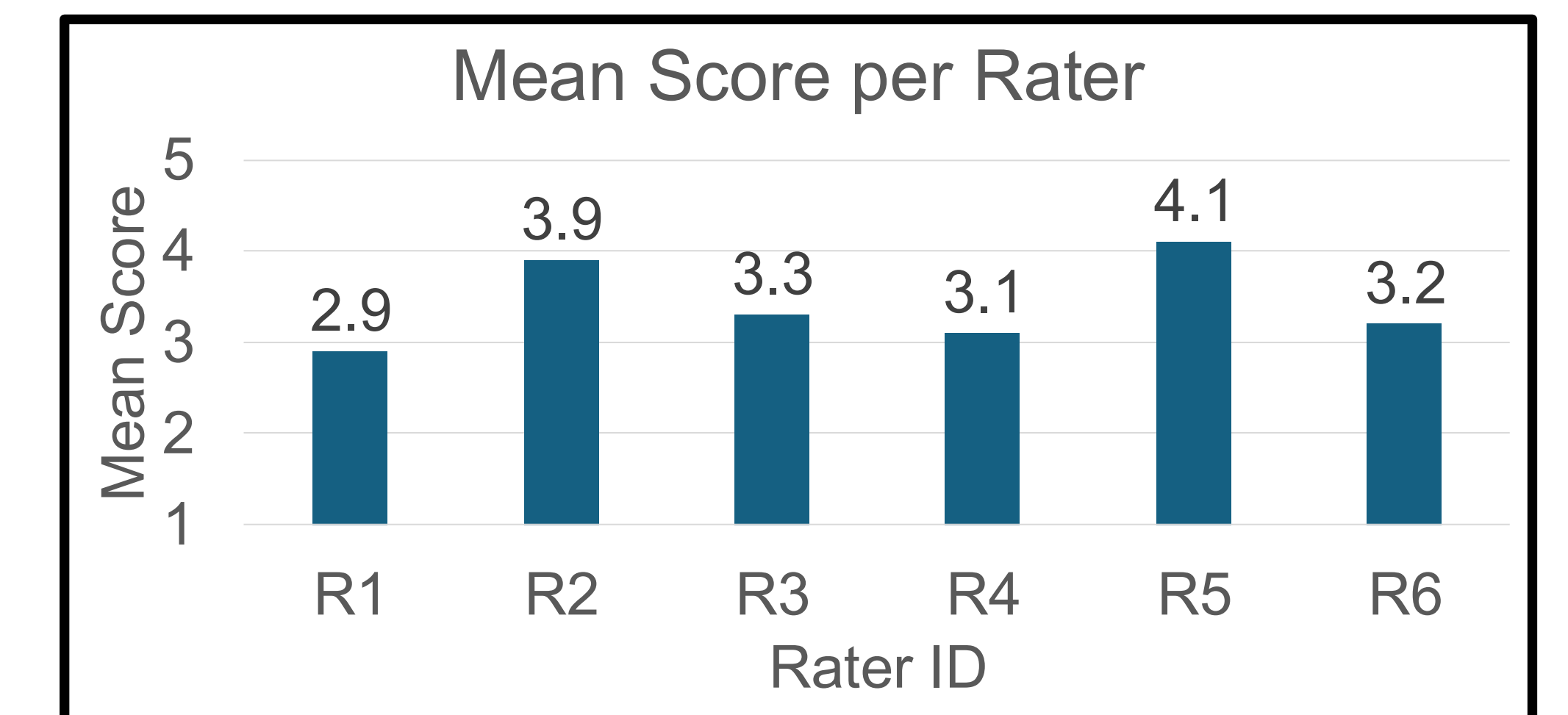
RESULTS

- A total of 8 videos were obtained between 6/2025 and 7/2025.
- Reviews were completed by 6 attending surgeons.
- The mean total modified GOALS score for novice (PGY1 through 3) was 12.8 (95% CI 10.74 to 14.93) compared with 14.3 (95% CI 11.60 to 16.90) for experienced residents (PGY4 through 5).
- Modified GOALS scores were positively correlated with resident PGY year (Spearman's $\rho = 0.59$, $p = 0.12$).
- Interrater reliability among the six attending surgeons was assessed using pairwise quadratic-weighted kappa.
- Agreement ranged from poor (R1-R2, $\kappa = -0.02$) to moderate (R3-R5, $\kappa = 0.45$). The overall mean kappa across all rater pairs was 0.18.
- Overall accuracy of reviewers predicting actual PGY level of training was 31.3% (95% CI Wilson 18.1% – 44.4%), with the PGY4 level being most accurately predicted and the PGY3 level being least accurately predicted.



Rater Pair	Quadratic Weighted Kappa (κ_w)	Interpretation
R3 vs. R5	0.45	Moderate Agreement (Benchmark)
R3 vs. R6	0.43	Moderate Agreement
R1 vs. R6	0.28	Fair Agreement
R2 vs. R5	0.22	Fair Agreement
R1 vs. R2	-0.02	No Agreement (Maximum Discrepancy)

- The "Hawk-Dove" effect is seen in the near-zero agreement between R1 and R2.
- The Hawk (R1) has the lowest mean score and the Doves (R2 and R5) have higher mean scores.



CONCLUSION

- Video-based assessment using objective rating scales can be used to reliably differentiate between expert and novice performance during laparoscopic appendectomy.
- Low agreement between raters and poor correlation of performance with PGY level of training indicates a need for standardization amongst reviewers.
- Even with a standardized rubric for evaluating laparoscopic appendectomy, subjective internal thresholds for competence vary substantially. Residents may experience inconsistent feedback and variable entrustability depending on attendings' individual scoring threshold rather than the residents' technical proficiency.
- Further study with comparison of intra-operative assessment with video assessment would add more meaningful data to help achieve competency benchmarks for each level of training.