ACS 2025 Surgeons and Engineers: A Dialogue on Surgical Simulation Meeting

P-E-05

Research Abstracts

Correcting for Rater Effects in the Assessment of Operative Difficulty and Surgical Skill in the Operating Room

Ryan Chou; Hajira Naz, BDS; Kofi D. O. Boahene, MD; Jessica H. Maxwell, MD; Ira D. Papel, MD; Theda C. Kontis, MD; Lisa E. Ishii, MD; Sonya Malekzadeh, MD, Masaru Ishii, MD, PhD; and S. Swaroop Vedula, MBBS, PhD, MPH

Whiting School of Engineering, Johns Hopkins University, Baltimore, MD; Dugoni School of Dentistry, University of Pacific, San Francisco, CA; Johns Hopkins University School of Medicine, Baltimore, Baltimore, MD; MedStar Georgetown University Hospital, Washington, DC

Introduction: Surgical skill assessment is integral to training, but its reliability and accuracy is reduced by rater effects and operative difficulty. Rater effects stem from a rater's strictness/leniency, ability, and stochastic variation. Operative difficulty can also be perceived differently by raters, even expert surgeons, and impacts assessment. Our objective in this study is to develop an accurate and reliable measure of surgical skill by adjusting for rater effects and operative difficulty.

Methods: Attending surgeons assessed residents and fellows performing nasal septoplasty in a prospective cohort study. We fit a structural equation model with Septoplasty Global Assessment Tool (SGAT) rubric item scores regressed on two rater-effect adjusted constructs: skill and operative difficulty, estimated from technical complexity, septal deflection and severity of pathology in each patient. We validated this model against the attending-assessed level of expertise and post-graduation year (PGY) commensurate with the trainee's performance, the actual PGY of the trainee, and achievement of surgical goals.

Results: Seven raters completed 188 assessments of 41 trainees. Five of seven SGAT items had significant operative difficulty coefficients, and all the rater coefficients for adjusting skill and operative difficulty were significant. The rater effect for skill had poor correlation with that for operative difficulty. Difficulty adjustment affected skill independent of rater adjustment (figure). Difficulty-adjusted latent skill scores increased with attending-estimated skill levels and PGY of trainees, increased with the actual PGY, and did not change significantly over different levels of achievement of surgical goals.

Conclusions: Our findings show a method to obtain a reliable and valid measure of surgical skill by simultaneously accounting for rater effects and operative difficulty. The rater effects for skill and operative difficulty should be estimated separately. Our method is necessary for reliable and useful national datasets used to set benchmarks for surgical skill.

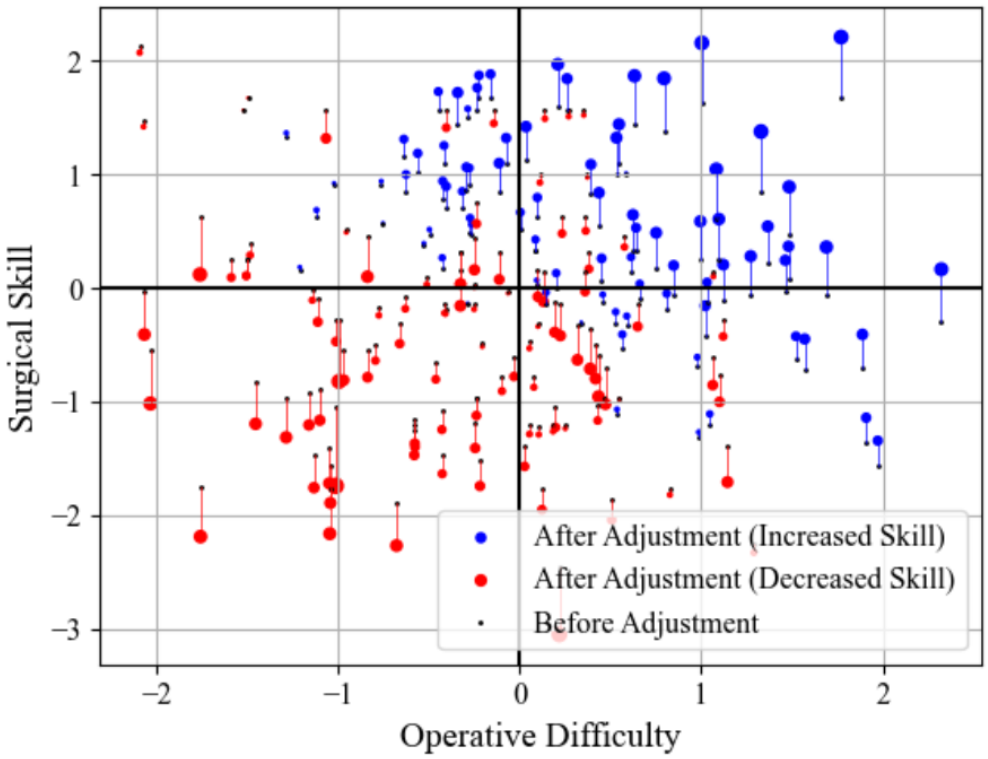

Figure: Adjusting for Operative Difficulty Changes Skill Scores
The black dots show skill scores adjusted only for rater effects. The red dots show skill scores that decreased after adjustment for both rater effects and operative difficulty. The blue dots show skill scores that increased after adjustment for both rater effects and operative difficulty.