A Preliminary Study on Crowdsourcing for Intraoperative Surgical Skill Assessment in Capsulorhexis

S. Swaroop Vedula, MBBS, PhD; Lauren Fang; Avigyan Sinha; Apurv H Shekhar; Austin Reiter, PhD;
Gregory D Hager, PhD; Shameema Sikder, MD*

Department of Computer Science, Whitng School of Engineering, Johns Hopkins University, Baltimore, Maryland, USA
*Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

WILMER EYE INSTITUTE

JOHNS HOPKINS UNIVERSITY WHITING SCHOOL OF ENGINEERING

JOHNS HOPKINS MEDICINE SCHOOL OF MEDICINE

Financial Disclosures:

- None

Synopsis:

- Crowdsourcing yields accurate assessment of surgical technical skill. Our work demonstrated interchangeability and validity of crowd ratings for intraoperative technical skill during capsulorhexis relative to expert assessments.

Introduction

- Cataract is an index surgery for residency training
- Residents expected to be competent by end of training
- ACGME has mandated objective evaluation

Current Assessment Tools

Table 1. Validity and reliability findings for surgical assessment tools.

Assessment Tool	Validity				Objective/Subjective	Reliability
	Face	Content	Construct	Predictive		
OASIS[1]	+	+	–	TBD	Objective	No inter-rater variability
GRASIS[2]	+	+	–	TBD	Subjective	TBD
HRACS[3]	+	+	+	TBD	Objective	TBD
OSCAR[4]	+	+	–	TBD	Objective	TBD
OSACSS[5]	–	–	+	TBD	Objective	TBD

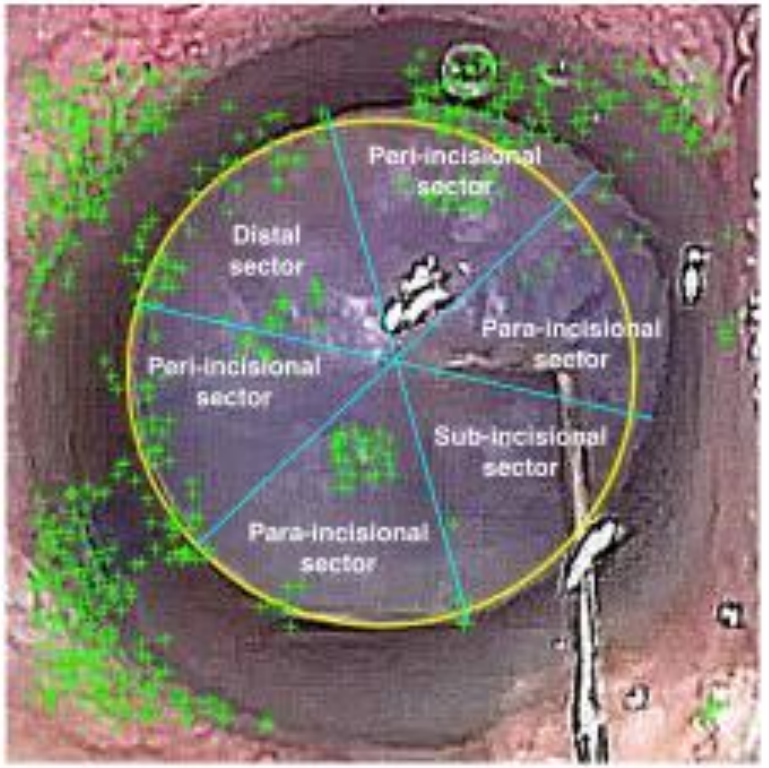
GRASIS = Global Rating Assessment of Skills in Intraocular Surgery; HRACS = Human Reliability Analysis of Cataract Surgery; OASIS = Objective Assessment of Skills in Intraocular Surgery; OSACSS = Objective Structured Assessment of Cataract Surgical Skill; OSCAR = Ophthalmology Surgical Competency Assessment Rubric; TBD = to be determined

- All dependent on real time observation
- Have elements of subjective assessment

Puri S, Sikder S, JCRS 2014

What is the next step in assessment?

- Technology to provide data-driven insights into surgical performance can automate summative assessments and feedback



What is crowd sourcing?

- An efficient method to obtain valid assessments of technical skill.
- Crowdsourcing technical skill assessment involves capturing evaluations from a large number of independent individuals who are not required to have expertise in surgery.
- Crowdsourcing has been shown to yield valid surgical technical skill assessments using global rating scales for several surgical tasks performed on bench-top simulation, animal models, and prostatectomy.

Chen J Surg Res 2014
Lendvay JAMA Surg 2015
Holst J Urol 2015

Objective

- Establish the reliability and validity of surgical technical skill assessments by a collective of surgically untrained individuals (i.e., a crowd) for capsulorhexis.
- We studied capsulorhexis because it is one of the most difficult aspects of cataract surgery to master.

Methods

- Captured video from operating microscope
- Conducted a survey of faculty surgeons (experts) and a collective of surgically untrained individuals (crowd) in a study approved by the Johns Hopkins Institutional Review Board.
- Faculty: 14 videos
- Residents: 13 PGY3 and 14 PGY4

- The crowd respondents were not educated about the surgical aspects of capsulorhexis as part of this survey other than being informed that a circular and regular tear in the capsule is considered an optimal end product.
- Survey:
  - capsulorhexis component in OSCAR and OSACSS
  - questions on circularity and overall performance of the capsulorhexis with both rated on a 5-point Likert scale
  - question on competency of the surgeon performing the capsulorhexis (not competent, competent to perform under supervision, competent to perform without supervision)
  - question on whether the operating surgeon was a faculty or trainee.

Results

Table 1. Intraclass correlations (ICC; 2,1) for ratings within experts and crowd

Survey item	ICC (95% confidence interval) for expert ratings	ICC (95% confidence interval) for crowd ratings
OSCAR – commencement of flap & follow-through	0.45 (0.31 to 0.60)	0.29 (0.17 to 0.45)
OSCAR – formation and circular completion	0.40 (0.26 to 0.55)	0.44 (0.31 to 0.59)
OSACSS – commencement of flap & follow-through	0.32 (0.19 to 0.48)	0.33 (0.20 to 0.48)
OSACSS – formation and circular completion	0.39 (0.25 to 0.54)	0.43 (0.30 to 0.58)
Circularity	0.41 (0.27 to 0.56)	0.38 (0.25 to 0.54)
Overall performance	0.36 (0.23 to 0.51)	0.32 (0.20 to 0.48)

The reliability within crowd ratings appears similar to that within expert ratings for all items in our survey except for OSCAR – commencement of flap & follow-through.

Results

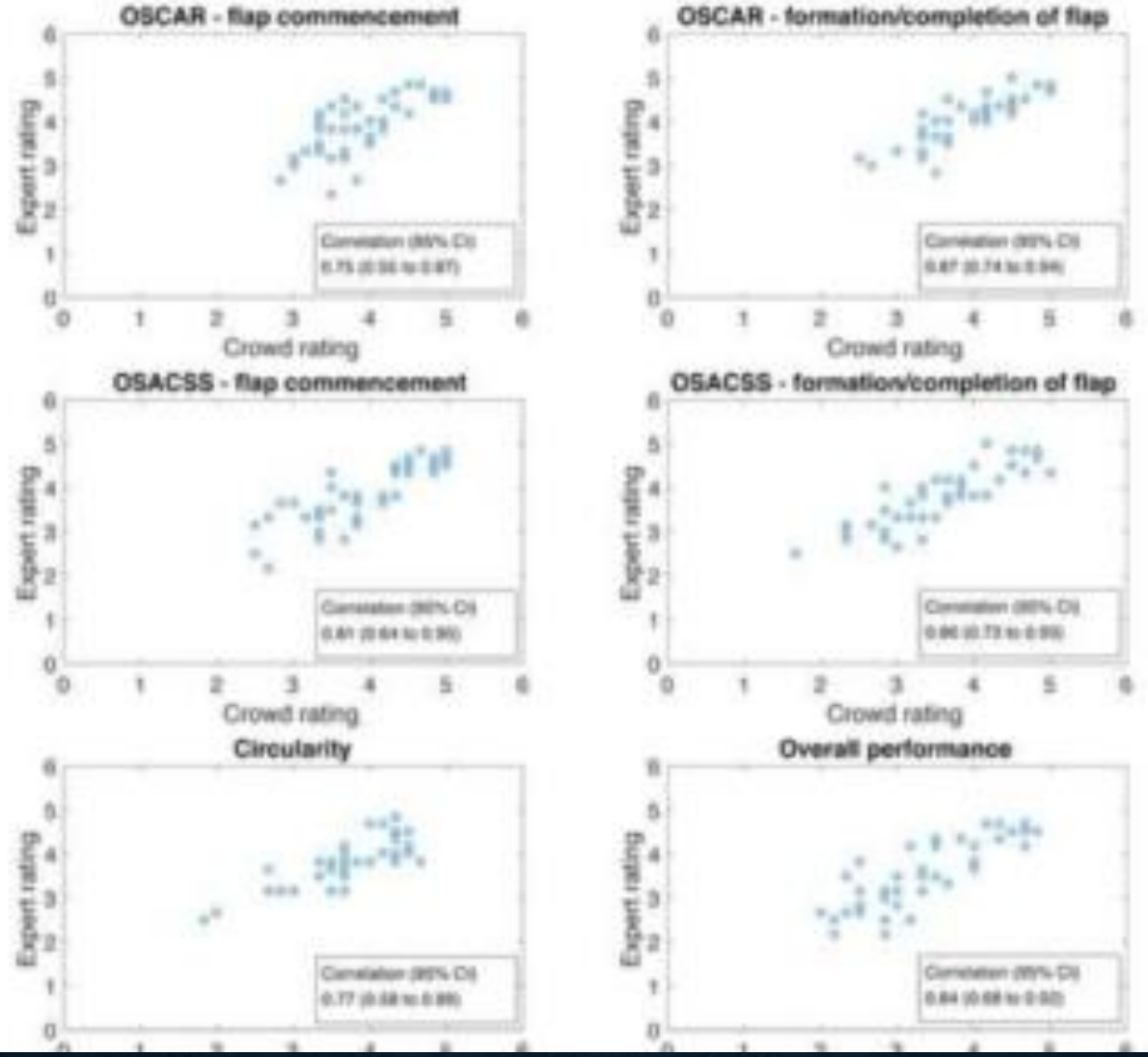
- There was good agreement between expert and crowd ratings for all survey items.
- The absolute difference in ratings was less than 0.5 for all items.
- The bias, which is the average of the actual difference in ratings, was small for questions on commencement of flap & follow-through but it was larger for the remaining items (Table 2).
- The LA indicate that crowd ratings for new instances of capsulorhexis videos will be approximately within one unit of expert ratings and the bias is equally likely in either direction.

Table 2. Limits of agreement between crowd and expert ratings (parentheses include 95% confidence intervals)

Survey item	Bias	Lower limit of agreement	Upper limit of agreement
OSCAR – commencement of flap & follow-through	-0.013 (-0.16 to 0.14)	-0.94 (-0.68 to -0.94)	0.92 (0.66 to 1.17)
OSCAR – formation and circular completion	0.125 (0.02 to 0.23)	-0.54 (-0.36 to -0.73)	0.79 (0.61 to 0.98)
OSACSS – commencement of flap & follow-through	-0.021 (-0.15 to 0.11)	-0.86 (-0.63 to -1.10)	0.82 (0.59 to 1.05)
OSACSS – formation and circular completion	0.225 (0.10 to 0.36)	-0.60 (-0.37 to -0.82)	1.05 (0.82 to 1.27)
Circularity	0.12 (-0.01 to 0.24)	-0.64 (-0.44 to -0.86)	0.88 (0.67 to 1.09)
Overall performance	0.17 (0.02 to 0.32)	-0.76 (-0.50 to -1.01)	1.10 (0.84 to 1.36)

The expected absolute difference between expert and crowd ratings is within one unit for all items in our survey except for OSACSS – formation and circular completion.

Correlation between expert and crowd ratings of surgical technical skill for capsulorhexis



- Crowd ratings were highly correlated with expert ratings for all items ($P < 0.01$; Figure 1), suggesting good criterion validity. The crowd was moderately accurate in assessing competence of the operating surgeon (accuracy = 0.75), and in predicting whether the operating surgeon was an attending or a trainee (accuracy = 0.8).

- Finally, experts and the crowd did not differ in the mean time they spent to view and rate each capsulorhexis video ($P = 0.8$). Experts spent 306 seconds on average (range = 110 to 709) and the crowd spent 332 seconds (range = 56 to 1173).

Discussion

- Our findings demonstrated that technical skill assessments for capsulorhexis by a surgically untrained crowd were reliable, interchangeable, and valid relative to assessments by expert surgeons.
- Crowdsourcing has previously been shown to yield valid assessments of surgical technical skill, and our study extends this previous work to skill assessment in the operating room in ophthalmology.
- The agreement we observed between expert and crowd ratings should be considered acceptable in a clinical context. The inter-rater reliability we observed in our study is consistent with previously published observations about reliability of expert ratings on OSCAR and OSACSS.

- While our study included a small crowd, their assessments were independent of each other.
- Our findings should be further validated in a subsequent study that involves a larger crowd.
- Our study focused on one of several tasks in cataract surgery. Evaluating skill for individual tasks in a procedure may provide more effective targeted feedback than an overall global score.

Conclusions

- We demonstrated reliability and validity of intraoperative technical skill ratings for capsulorhexis by a crowd.
- Our findings support further studies to verify our preliminary observations on a large scale, and on integration of crowdsourced technical skill assessment using validated structured scales into residency training curricula.

Acknowledgements

- Anand Malpani provided code for the web interface and critically reviewed our findings and earlier versions of this manuscript.
- We also acknowledge contributions by survey respondents.

Research to Prevent Blindness