

# A Competency Assessment Tool for Unilateral Cleft Lip Repair

Carolyn R. Rogers-Vizena, MD\*<sup>†</sup>  
 Georgios D. Sideridis, PhD<sup>††</sup>  
 Krishna G. Patel, MD<sup>§</sup>  
 Catharine B. Garland, MD<sup>¶</sup>  
 Delora L. Mount, MD<sup>¶</sup>  
 Caroline A. Yao, MD<sup>||\*\*</sup>

**Background:** Objective evaluation of operative performance is increasingly important in surgical training. Evaluation tools include global rating scales of performance and procedure-specific skills checklists. For unilateral cleft lip repair, the numerous techniques make universal evaluation challenging. Thus, we sought to create a unilateral cleft lip evaluation tool agnostic to specific repair technique.

**Methods:** Four surgeons with expertise in 3 common cleft lip repair techniques participated in a 3-round Delphi process to generate consensus evaluation points spanning all techniques. Items were categorized as marking the repair, performing the repair, and final result. Two blinded raters then scored videos of simulated cleft lip repairs using both the 21-item novel checklist and the modified Objective Structured Assessment of Technical Skills. Kappa and *T* values were calculated for both scales to determine level of agreement.

**Results:** Ten videos of repairs performed by novice residents through experienced craniofacial fellows were scored. Moderate ( $\kappa = 0.41$ – $0.60$ ) to substantial ( $\kappa = 0.61$ – $0.80$ ) interrater reliability was seen for the majority of questions in both the novel tool and the Objective Structured Assessment of Technical Skills. A single question in the novel tool had almost perfect agreement ( $\kappa = 0.81$ – $1.00$ ), 8 had moderate agreement, and 6 had substantial agreement. Poorly scoring questions were discarded from the final 18-item tool.

**Conclusions:** Despite variations in unilateral cleft lip repair technique, common themes exist that can be used to assess performance and outcome. A universal evaluation tool has potential implications for trainee assessment, surgeon credentialing, and screening for surgical missions. (*Plast Reconstr Surg Glob Open* 2020;8:e2954; doi: 10.1097/GOX.0000000000002954; Published online 14 July 2020.)

## INTRODUCTION

Objective performance evaluation is increasingly important for surgical trainee evaluation and is emerging as a component of credentialing. Evaluation tools include global rating scales of performance, most commonly the Objective Structured Assessment of Technical Skills,<sup>1</sup> and

procedure-specific skills checklists pertaining to an individual operation.

Training surgeons to perform cleft repair has been difficult, given disease rarity and the high stakes of operating on an infant's face. Because unilateral cleft lip repair is a core procedure for plastic, oral, and otolaryngology surgical training, a procedure-specific checklist would aid in evaluating performance and in determining readiness for practice. Furthermore, there is value in the surgical mission setting, where verifying participant competency and surgical outcome is increasingly important. However, structured evaluation of unilateral cleft lip repair is

From the \*Department of Plastic and Oral Surgery, Boston Children's Hospital, Boston, Mass.; †Harvard Medical School, Boston, Mass.; ‡Institutional Centers for Clinical and Translational Research, Boston Children's Hospital, Boston, Mass.; §Department of Otolaryngology, Medical University of South Carolina, Charleston, S.C.; ¶Division of Plastic and Reconstructive Surgery, University of Wisconsin Hospital and Clinics, Madison, Wis.; ||Shriners for Children Medical Center—Pasadena, Pasadena, Calif.; and \*\*Operation Smile, Virginia Beach, Va.

Received for publication March 26, 2020; accepted May 8, 2020.

Copyright © 2020 The Authors. Published by Wolters Kluwer Health, Inc. on behalf of The American Society of Plastic Surgeons. This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

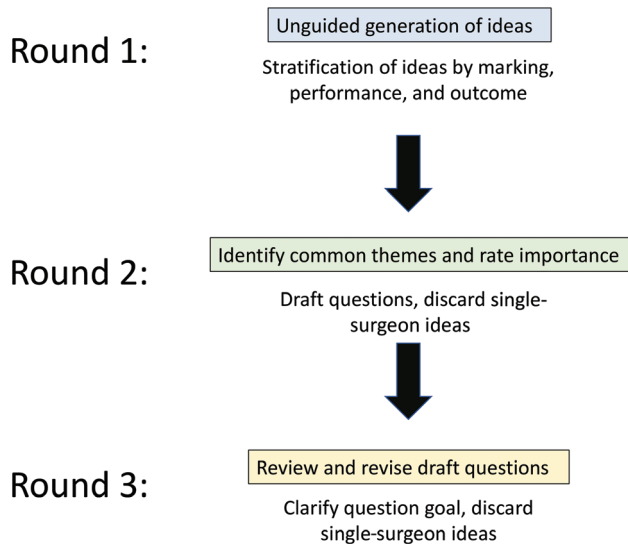
DOI: 10.1097/GOX.0000000000002954

**Disclosure:** The authors have no financial interest to declare in relation to the content of this article.



This work was supported by  
THE PLASTIC SURGERY FOUNDATION.

Related Digital Media are available in the full-text version of the article on [www.PRSGlobalOpen.com](http://www.PRSGlobalOpen.com).



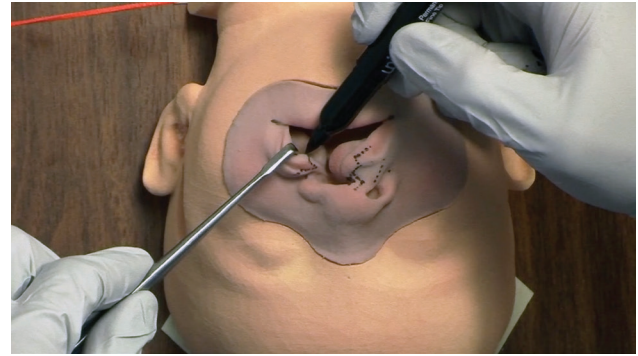
**Fig. 1.** Delphi process for developing the unilateral cleft lip assessment tool. Round 1: Unguided generation of ideas, followed by stratification of ideas into core domains—marking, performance, and final outcome. Round 2: Review of aggregate ideas to identify common themes that spanned technique (eg, “Reposition the greater segment inferiorly so peaks of Cupid’s bow are symmetric”), followed by drafting of specific data points and discarding either ideas supported by only one surgeon or those specific to only one technique. Round 3: Review of draft data points, followed by revision to clarify goal of the data point and discarding either points endorsed by only one surgeon or those specific to only one technique.

challenging due to existence of numerous techniques without consensus on which is best. We sought to create an evaluation tool agnostic to specific technique to allow broad-based performance evaluation.

## METHODS

This work was completed under institutional review board exemption status. A panel of 4 surgeons was assembled with expertise in the 3 most common eponymous unilateral cleft lip repairs: Millard, Mohler, and Fisher. A 3-round Delphi process was undertaken to formulate consensus about the most important aspects of repair spanning all techniques (Fig. 1). Although all 4 surgeons perform primary rhinoplasty and nasal repair evaluation was considered, it was not included in the final tool because primary rhinoplasty is not universally performed. After a consensus was reached, a 19-item checklist was generated addressing the following: marking the repair, performing the procedure, and final result. Questions were reviewed by a survey methodologist to minimize bias and to develop a 3-pronged assessment scale.

To appraise validity, videos of operators of varying training levels performing a unilateral cleft lip repair of their choice on identical surgical trainers<sup>2</sup> were used. The trainer was used so all participants, regardless of skill level, could perform the procedure uncoached from start to finish to determine the tool’s ability to discriminate across the performance spectrum while standardizing cleft severity. Videos were recorded from a single closeup



**Fig. 2.** Example of video setup. Unilateral cleft lip repairs were completed on a surgical simulator and recorded from a full-face frontal view. Sound was removed from videos for the purpose of anonymizing the operator for blinded rating.

frontal view (Fig. 2) without sound. The 2 raters blindly scored all recordings using both our novel tool and the abbreviated OSATS scale<sup>3</sup> (Table 1). One rater was from the initial instrument development group and 1 was not involved in development and thus worked directly from the language of the survey unbiased by the planned goals of the group.

After preliminary rating, we determined that although the scale adequately assessed intended goals of repair, it did not account for problems of design and execution unique to surgeons in training. Questions addressing abnormally retaining or resecting tissue and abnormal dissection were added to account for these issues (Table 2). The videos were re-rated. Items were scored as follows: 1, performed incorrectly or not at all; 2, performed somewhat correctly; 3, performed correctly, acceptable for an attending surgeon. Interrater reliability was assessed for both the novel checklist and OSATS. Reliability was reported using kappa and T value. Kappa results were differentiated using criteria proposed by Cohen<sup>4</sup>: values  $\leq 0$  indicating no agreement, 0.01–0.20, none to slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, almost perfect agreement. Significance was determined as  $P < 0.05$ .

## RESULTS

Videos were captured from 10 PGY 3–8 plastic surgery residents and craniofacial fellows. Intended style of repair was evident for 5 (4 classic extended-Mohler and 1 classic Mulliken-modification rotation advancement). The other 5 did not adhere to classic techniques and were best categorized as 2 rotation advancement–like repairs, 2 Mohler-like repairs, and 1 triangular flap variant. Kappa values for interrater reliability revealed that of the 21 items evaluated, 1 had almost perfect agreement ( $\kappa = 1.000$ ), 6 had substantial agreement ( $\kappa$  ranges, 0.623–0.800), 8 had moderate agreement ( $\kappa$  ranges, 0.429–0.600), 6 had fair agreement ( $\kappa$  ranges, 0.265–0.400), and no questions had no or only slight agreement (Table 2). Similar interrater reliability results were seen for OSATS, where out of 4 questions, 2 had substantial agreement ( $\kappa$  ranges,

**Table 1. Modified OSATS Global Rating Scale<sup>3</sup>**

Items	Kappa Coefficient	T Value	Agreement Level
Respect for tissue 1 = Frequently used unnecessary force on tissue or caused damage by inappropriate use of instruments 3 = Careful handling of tissue but occasionally caused inadvertent damage 5 = Consistently handled tissue appropriately with minimal damage	0.437	2.450*	Moderate
Time and motion 1 = Many unnecessary moves 3 = Efficient time/motion but some unnecessary moves 5 = Economy of movement and maximum efficiency	0.683	3.064*	Substantial
Instrument handling 1 = Repeatedly makes tentative or awkward moves with instruments 3 = Competent use of instruments although occasionally appeared stiff or awkward 5 = Fluid moves with instruments and no awkwardness	0.394	1.840†	Fair
Flow of operation and forward planning 1 = Frequently stopped operating or needed to discuss next move 3 = Demonstrated ability for forward planning with steady progression of operative procedure 5 = Obviously planned course of operation with effortless flow from one move to the next	0.722	3.893*	Substantial

\**P* < 0.05 using a 1-tailed test.†*P* < 0.05 using a 1-tailed test.**Table 2. Interrater Reliability of Initial Items for the Unilateral Cleft Lip Repair Assessment Tool**

Items	Kappa Coefficient	T Value	Agreement Level	Recommendation
Marking a cleft lip repair				
Correctly identify anatomic landmarks	0.800	2.582*	Substantial	Keep
Mark appropriate peaks of Cupid's bow on the greater and lesser segments	0.756	3.266*	Substantial	Keep
Design greater segment incision(s) to create a symmetric Cupid's bow	1.000	3.943*	Almost perfect	Keep
Preserve adequate skin and mucosa to construct a symmetric columella and nasal sill	0.375	1.779†	Fair	Discard
Mark lesser segment incisions to create a symmetric philtral column	0.400	1.291	Fair	Discard
Preserve excess vermilion on the lesser segment to augment deficient vermilion on the greater segment	0.583	1.845†	Moderate	Keep
Performing a cleft lip repair				
Precise and accurate incisions	0.286	1.291	Fair	Discard
Atraumatic tissue handling	0.600	2.070*	Moderate	Keep
Avoid unnecessarily retaining or overresecting tissue	0.531	2.348*	Moderate	Keep
Dissect orbicularis oris muscle and free its abnormal attachments	0.677	2.896*	Substantial	Keep
Inferiorly reposition greater segment so the peaks of Cupid's bow are symmetric	0.524	2.328*	Moderate	Keep
Fully mobilize the lesser segment lip and alar base	0.524	2.741*	Moderate	Keep
Avoid overdissection or underdissection	0.667	2.843*	Substantial	Keep
Repair the oral mucosa	0.355	1.517	Fair	Further evaluation needed
Repair the orbicularis oris muscle avoiding deficiency or excess bulk	0.692	3.116*	Substantial	Keep
Close the nasal floor mucosa	0.545	2.463*	Moderate	Keep
Appropriate suture choice and precise cutaneous repair to minimize scarring	0.552	2.782*	Moderate	Keep
Immediate postoperative result				
Symmetric appearance of the peaks of Cupid's bow and philtral height	0.265	1.363	Fair	Further evaluation needed
Continuity of the vermilion–cutaneous junction and white roll	0.429	2.673*	Moderate	Keep
Smooth contour of the lip margin	0.286	1.243	Fair	Further evaluation needed
Symmetry of the nasal tip, alar position, and nares	0.623	2.923*	Substantial	Keep

Grading scale for all question: 1, performed incorrectly or not at all; 2, performed somewhat correctly; 3, performed correctly, acceptable for an attending surgeon.

\**P* < 0.05 using a 1-tailed test.†*P* < 0.05 using a 1-tailed test.

0.683–0.722), 1 had moderate agreement ( $\kappa = 0.437$ ), and 1 had fair agreement ( $\kappa = 0.394$ ; [Table 1](#)).

## DISCUSSION

Structured evaluation is an increasingly important part of surgical training, as evidenced by the Accreditation Council for Graduate Medical Education's growing emphasis on Milestones.<sup>5</sup> It adds unique assessment beyond traditional tools like in-service or board exams. Although testing for core knowledge is important, performance and technical skill is a key element of competency. The last several

decades have brought interest in structured, validated tools for assessing operative performance. Both global rating scales<sup>6</sup> and procedure-specific checklists for appraising critical steps of an individual operation continue to be developed. With growing interest in competency-based resident graduation,<sup>7</sup> increased availability of cleft simulators for training and evaluation, and growing scrutiny of outcomes<sup>8</sup> and educational impact<sup>9</sup> of cleft missions, there is need for a tool to evaluate cleft lip repair competency.

We demonstrate that despite variation in technique, common unilateral cleft lip repair themes can be used to

assess competency. As a whole, our proposed procedure-specific checklist demonstrates comparable interrater reliability to OSATS, with the majority of questions having moderate to almost-perfect agreement. For 1 question having only fair agreement, specifically “repair the oral mucosa,” camera angle may have limited intraoral visualization, forcing raters to speculate about execution. This could be resolved by scoring in person, so that the rater can assess performance from different angles. Consequently, discarding this question is premature, but further scrutiny is needed.

We recommend discarding a few items with only fair reliability to improve reproducibility of the scale. “Mark lesser segment incisions to create a symmetric philtral column” may have been unreliable because some operators trim as needed rather than marking the incision. As surgeons who mark precise incisions and those who “cut as you go” can both have excellent results, we propose discard. Although we initially assumed the frontal view limited “preserve adequate skin and mucosa to construct a symmetric columella and nasal sill,” one author recently began to apply this tool with in-person simulation and continued to have difficulty rating; therefore, we recommend discard. A surprise was the low reliability for “precise and accurate incisions.” The compound statement may be the source of discrepancy, as incisions can be “precise” but not “accurate.” We recommend discarding this item.

Postoperative result items performed inconsistently. This may speak of differing philosophies about what an “acceptable” repair should look like. Consequently, other validated pictorial scales for postoperative result could be substituted<sup>10</sup> to minimize bias. Moreover, it drives home the importance of quantitative measurement to provide objective outcome assessment. The final recommended tool is found in the unilateral cleft lip repair procedural checklist (see **Checklist, Supplemental Digital Content 1**, which displays the final version of the unilateral cleft lip repair assessment tool, <http://links.lww.com/PRSGO/B421>).

Because there are advantages to both procedure-specific checklists and global rating scales, we advocate use of both as part of a holistic approach to structured evaluation. They can be used on clinical rotations to evaluate resident performance or to determine readiness for independent practice. Structured, validated tools could also be used in board certification to add a technical element to credentialing. A final implication is for humanitarian missions, where this tool paired with simulation offers opportunity to screen competency permission and for ongoing scrutiny of surgical work.

### LIMITATIONS

Validation in different contexts with multiple raters is needed, particularly in-person evaluation and with different cleft morphology. Because the tool is technique-agnostic, it does not capture subtleties of each eponymous repair. Furthermore, having agnostic raters is paramount

to accuracy, as many surgeons have strong preferences and may find it difficult to objectively evaluate a technique they do not use. Finally, this tool cannot be applied to other operations. However, the strategy of multiround Delphi process to formulate consensus questions followed by blinded rating could serve as a framework for developing other procedural checklists.

**Carolyn R. Rogers-Vizena, MD**

Department of Plastic and Oral Surgery  
Boston Children’s Hospital  
300 Longwood Ave  
Boston, MA 02115

E-mail: [Carolyn.rogers@childrens.harvard.edu](mailto:Carolyn.rogers@childrens.harvard.edu)

### ACKNOWLEDGMENTS

*This research is part of a broader project funded by a National Endowment for Plastic Surgery grant through The Plastic Surgery Foundation. The simulator used in this study is proprietary technology developed by the first author in conjunction with the Boston Children’s Hospital Simulator Program. This technology is currently not publicly available. This work was slated for presentation at the annual meeting of the American Cleft Palate-Craniofacial Association, Portland, Ore., April 1, 2020, but has been postponed due to the current COVID-19 pandemic.*

### REFERENCES

1. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84:273–278.
2. Rogers-Vizena CR, Saldanha FYL, Hosmer AL, et al. A new paradigm in cleft lip procedural excellence: creation and preliminary digital validation of a lifelike simulator. *Plast Reconstr Surg*. 2018;142:1300–1304.
3. C-SATS. Objective Structured Assessment of Technical Skills (OSATS). Johnson & Johnson Co. Available at <https://www.csats.com/osats/>. Accessed February 11, 2020.
4. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37–46.
5. The Accreditation Council for Graduate Medical Education (ACGME). Surgery Milestones, 2nd revision. 2019. Available at <https://www.acgme.org/Portals/0/PDFs/Milestones/SurgeryMilestones2.0.pdf?ver=2019-05-29-124604-347>. Accessed March 3, 2020.
6. Reznick R, Regehr G, MacRae H, et al. Testing technical skill via an innovative “bench station” examination. *Am J Surg*. 1997;173:226–230.
7. Nguyen VT, Losee JE. Time- versus competency-based residency training. *Plast Reconstr Surg*. 2016;138:527–531.
8. Daniels KM, Yang Yu E, Maine RG, et al. Palatal fistula risk after primary palatoplasty: a retrospective comparison of a humanitarian organization and tertiary hospitals. *Cleft Palate Craniofac J*. 2018;55:807–813.
9. McCullough M, Campbell A, Siu A, et al. Competency-based education in low resource settings: development of a novel surgical training program. *World J Surg*. 2018;42:646–651.
10. Campbell A, Restrepo C, Deshpande G, et al. Validation of a unilateral cleft lip surgical outcomes evaluation scale for surgeons and laypersons. *Plast Reconstr Surg Glob Open*. 2017;5:e1472.