Psychometric Properties of a New Tool to Assess Task-Specific and Global Competency in Cataract Surgery

Dan B. Rootman, MSc, MD; Kay Lam, MD; Marisa Sit, MD, FRCSC; Eugene Liu, MD, FRCSC; Adam Dubrowski, PhD; Wai-Ching Lam, MD, FRCSC

BACKGROUND AND OBJECTIVE: To establish and validate an assessment tool of cataract surgery performed by residents suitable for a competency-based curriculum.

PATIENTS AND METHODS: A three-component evaluation tool was created based on review of the literature and was refined using a modified Delphi technique. Faculty surgeons viewed two videos of cataract surgery, performed by a novice and an expert, and completed the evaluation tool. Results were analyzed for the psychometric properties.

RESULTS: Evaluators concluded the scale had excellent face validity. Construct validity showed the scale to reliably distinguish \( P < .001 \) between novice (30.3 ± 6.1) and experienced (48.3 ± 7.2) surgeons. Internal consistency of the scale was high, with Cronbach’s alpha equal to 0.981. Inter-rater reliability was high with an intraclass correlation coefficient equal to 0.811 \( (F(df) = 53.2 (25), P < .001) \).

CONCLUSION: The tool has excellent face validity, content validity, and reliability. Its task-specific, global-index scale and quantitative data form make it a valuable tool to assess residents’ surgical skills.


INTRODUCTION

The competency-based accreditation model for surgical training asserts that trainees should achieve predefined levels of clinical and surgical competency to gain certification and ultimately provide safe and effective medical services to the public. This concept does not necessarily imply a predefined learning interval, but rather is focused on the achievement of specific competency goals.¹ To certify trainees in this compe-

From the University of Toronto (DBR, MS, KL, AD, W-CL), Toronto, Ontario, Canada; University Health Network (MS, W-CL), Toronto, Ontario, Canada; Mt. Sinai Hospital (EL), Toronto, Ontario, Canada; and the Wilson Centre (AD), University of Toronto, Toronto, Ontario, Canada.

Presented as a poster at The Royal College’s International Conference on Residency Education, September 23-25, 2010, Ottawa, Ontario, Canada.
The authors have no financial or proprietary interest in the materials presented herein.
Address correspondence to Dan B. Rootman, MSc, MD, Toronto Western Hospital, East Wing 6-401, 399 Bathurst Street, Toronto, Ontario M5T 2S8, Canada.
E-mail: dan.rootman@gmail.com
doi: 10.3928/15428877-20120315-02
tency-based model, valid and reliable assessments of surgical performance are required.

Currently, much of trainee surgical evaluation is based on global cumulative expert assessment of overall ability by surgeon supervisors in the form of an In-Training Evaluation Report. However, these summative reviews are dependent on preceptor recollection of surgical skills at the end of a rotation and tend to be subjective, vague, and unreliable. Procedure logs are also often employed, but these are only able to show quantity of experience and may not accurately reflect the quality of surgical proficiency. In response to these challenges, educators and accreditors alike have called for more direct measurements of surgical competencies. As specifically applied to ophthalmology, such a tool for the assessment of cataract surgery would be important to develop and use because this is one of the most commonly performed ophthalmic procedures.

When developing an assessment tool, it is important to consider that surgical competence itself is a complex quality that encompasses cognitive knowledge, technical skills, professional attitudes, and judgment in the context of real-life surgical scenarios. Any assessment of competency in phacoemulsification should consider all of these dimensions.

Current tools incorporate only partial components of this multidimensional process. Tools using global rating scores acknowledge that desirable global attributes would be recognized by expert surgeons, and have demonstrated validity and reliability. But these tools are not specific to a particular procedure. Alternatively, task-specific checklists that dissect a particular surgery into discrete elements are useful in procedure-specific teaching and additionally can improve the objectivity of the examiner. However, such checklists alone do not cover the global aspects of surgery and do not provide measures of competency in performing each step. Finally, quantitative data are also valuable to objectively assess surgical events and document complications. These measurements are important in defining the hard outcomes of surgery, particularly patient safety and anatomical success.

Combining these assessment strategies would create a tool that could maximize the advantages and minimize the disadvantages of each approach. Such a tool may more accurately assess the multidimensional quality of surgical learning.

In this investigation we aimed to create such an assessment tool, the Standardized Evaluation Protocol of Cataract Surgery (SEPOCS). This form combines task-specific and global-skills elements and a simple form for quantitative data collection. Our goal is to create a reliable, valid, and practical assessment tool to evaluate the cataract surgical competency of ophthalmology trainees. In this study, we assessed the face validity, reliability, and content validity of the SEPOCS scale.

**PATIENTS AND METHODS**

This prospective psychometrics study was approved by the University of Toronto Research Ethics Board and conformed to the tenets of the Declaration of Helsinki. The project was divided into two stages: the first focused on the establishment of face validity and the second concentrated on verifying the reliability and content validity of the scale.

**Face Validity**

The first stage of data collection followed a modified Delphi technique. A trial scale was created based on clinical experience and review of the literature. Some modifications were made to amalgamate features from these scales. Further amendments were made based on institutional experience with this amalgamated scale (unpublished data).

A web-based survey of high-volume primary cataract surgery teachers in the institution was performed. Each surgeon received the basic scale as an e-mail attachment and completed a standardized form assessing opinions in three main areas. The first section was overall scale ease-of-use and appropriateness. The second section used two 5-point Likert scales to assess the clarity and importance of each element in the task-specific and global indices. The last section used free-form comment entry to indicate any elements that were thought to be incompletely represented in the scale or the quantitative data form.

Any element of the scale that was deemed not to be “completely clear” by more than 70% of the respondents was modified for clarity. Any element that was not deemed to be “very important” by more than 80% of the respondents was removed. Any elements thought to be missing from the original data form were added to the new scale.

This modified scale was again e-mailed to the same surgery teachers and a similar online evaluation survey was performed. The same criteria for modification, ad-
dition, and removal of items were used on this second round of evaluation.

Reliability

For the reliability analysis, two videos were graded by cataract surgeons using the final SEPOCS scale. The first video was of an experienced surgeon (> 500 cases) and the second was of a novice surgeon (< 50 cases). Reviewers were blinded as to the identity of the surgeon in the video. Videos were viewed via the Internet and a web-based version of the SEPOCS scale was filled out for each. Some elements that were inappropriate for video-based review were removed from the scale (eg, interaction with operating room team).

Reliability analysis was based on all video reviews. Analyses included assessment of internal consistency with Cronbach's alpha. Inter-item and item total correlations were additionally calculated to assess the inter-relationship of individual elements. Factor analysis was performed with the intention that the scale should measure a single factor, that of surgical skill. Finally, intraclass correlation coefficient was calculated to assess inter-rater reliability.

Content validity was assessed for the task-specific, global, and total scores using the Student's t test to compare mean scores for the experienced and novice surgeon. All analyses were performed using SPSS 18.0 software for Mac (IBM SPSS Statistics, Armonk, NY).

RESULTS

Face Validity

Of the 20 cataract surgeons invited to participate in the process, 14 (70%) responded. The instructions for use of the scale were rated as self-explanatory by 100% of surgeons. Overall, the scale was rated to be a good measure of surgical performance by 92.9% of respondents.

During the first round, one item in the task-specific scale (hydrodelineation) was not considered completely clear or important. This item was removed for round two. Additionally, surgical pause, patient preparation, and draping were all noted as important components of the task-specific process and were added as a single element for round two. Several changes to the captions were made to enhance clarity.

Most items were not rated as completely clear in

| TABLE 1 |
|---|---|
| Results of Round 2 Modified Delphi Technique Face Validity Assessment | |
| | Completely Clear | Very Important |
| Task specific | | |
| Surgical pause, prep, drape, & speculum | 90.9% (10) | 81.8% (9) |
| Wound construction | 100% (11) | 100% (11) |
| Capsulorhexis | 90.9% (10) | 100% (11) |
| Hydrodissection | 90.9% (10) | 90.9% (10) |
| Phacoemulsification | 100% (11) | 100% (11) |
| Irrigation & aspiration | 100% (11) | 90.9% (10) |
| Lens placement | 100% (11) | 100% (11) |
| Wound closure | 100% (11) | 90.9% (10) |
| Global index | | |
| Microscope use & centration | 72.7% (8) | 90.9% (10) |
| Treatment of intraocular structures | 90.9% (10) | 100% (11) |
| Equipment handling & use of both hands | 90.9% (10) | 90.9% (10) |
| Management of unexpected events | 81.8% (9) | 90.9% (10) |
| Flow of operation | 72.7% (8) | 90.9% (10) |
| Interaction with operating room team | 72.7% (8) | 90.9% (10) |
| Case to case consistency | 90.9% (10) | 100% (11) |
the global scale and two items were not considered highly important. This scale was altered for clarity and the two items considered not highly important (instrument handling and use of non-dominant hand) were combined into a single element. In addition, interaction with the operating room team was raised as an important issue for the global scale and was added.

A few elements were added to the quantitative form during this process, including complications such as wound burn, conjunctival chemosis, detached Descemet’s membrane, vitreous loss, and iris prolapse. Case types were added, including monocular, mature: dense, mature: white, and soft. The participation list was modified to include anterior vitrectomy.

In round two, every respondent thought that, overall, the new scale was superior to the previous scale. Each element in the round two scale was rated as completely clear by more than 90% of respondents. The majority of elements were rated as very important by more than 90% of responders. Each element in the global scale was rated with improved clarity and also rated to be very important by more than 90% of surgeons. No items were added to the quantitative data form. All of the final results can be found in Table 1. The figure describes the final SEPOCS evaluation form.

### Reliability

Internal reliability of the scale was high, demonstrating a Cronbach’s alpha equal to 0.981. This number did not change significantly for the overall scale if each item was removed individually. Inter-item correlation was universally high, with no correlation below $r = 0.694$. Similarly, item-total correlations were high with none below $r = 0.817$. Factor analysis revealed a single component with eigenvalue greater than 1 (eigenvalue for factor 1 = 9.34), indicating the scale measures a single domain. Finally, inter-rater reliability was high with an intraclass correlation coefficient equal to 0.811 ($F(df) = 53.2 (25), P < .001$).

### Construct Validity

When scored by 14 expert observers, each subscale and the overall total score was able to reliably distinguish between novice and experienced surgeons. For the task-specific scale, the mean score for novice surgeons was 11.6 points less than for experienced surgeons ($t(df) = -7.05 (25), P < .001$). The same was true for the global index subscale (mean difference, $t(df) = -6.41, -6.51 (25), P < .001$) and the total score (mean difference, $t(df) = -18.0, -6.99 (25), P < .001$). These results are summarized in Table 2. The scale took a mean (standard deviation) of 3.75 (6.38) minutes to complete.

### DISCUSSION

Existing methods for evaluation of surgical skills including logbooks and non-criteria–based observations have demonstrated poor validity and reliability. Standardized tools such as the SEPOCS scale evaluated in this study aim to provide a more complete and valid form of assessment for competency in cataract surgery. The SEPOCS scale demonstrated excellent face
validity, content validity, and reliability. More than 90% of surgeons agreed that the SEPOCS’ task-specific global scales and quantitative data collection forms were appropriate for the assessment of resident surgical skills. The scale also demonstrated ability to distinguish trainee surgeons from expert surgeons on task-specific, global, and overall scores. Reliability metrics were high, demonstrating strong internal consistency and inter-rater reliability.

Some limitations of this study should be noted. Not every cataract surgery teacher in our institution decided to participate, thus some opinions and evaluations will be missing. However, a sample of 15 experts is traditionally thought to be sufficient for Delphi techniques in homogeneous populations (ie, cataract surgery teachers).8 Our sample of 14 is only 1 shy of this number. Additionally, we did not assess multiple trainee levels in the construct validity analysis. Such an analysis was beyond the primary purpose of this project, which was to establish the face validity and reliability of the scale, in addition to basic construct validity. A longitudinal analysis of individual trainee progression would be ideal to establish the subtle differentiating characteristics of the SEPOCS tool as learners improve with experience. The establishment of a reliable and valid measure to use, as performed here, would be a prerequisite for such an investigation.

The SEPOCS scale defined in this article combines task-specific (checklist + ratings), global, and quantitative elements, maximizing the overall utility of the tool by using multiple evaluation elements, each with their own advantages. Task-specific scales can be defined as a complete set of separate actions that constitute the performance of a particular task. By using explicit and predefined criteria of technical skill, the aim of such scales is to increase the objectivity of the evaluation process. Regehr et al. noted that checklists turn examiners into observers of behavior rather than interpreters of behavior, thereby removing the subjectivity of the evaluation process.5 Skills assessed with such measurements avoid the “halo effect” (tendency to judge favorably based on a few positive characteristics) and the “pitchfork effect” (tendency to judge unfavorably based on a few negative characteristics) in the scoring of the subjective global evaluation form.6 Furthermore, by separating specific actions, certain steps that are more difficult to master are emphasized. For instance, such rating can point a resident to focus practice on a particularly difficult skill such as capsulorhexis.9

However, the opportunity for expert raters to form a holistic impression of overall performance is lost if one relies only on a checklist. Previous studies have shown that global rating scores by expert examiners have enhanced validity and reliability when compared with the checklist scores alone.5 Expert surgeons are able to recognize desirable qualities even if they are more challenging to explicitly define. Thus, although global scales may assess generic surgical skills, they integrate expert experience into the assessment and provide a more holistic evaluation of competence.

Finally, although both checklists and global rating scales have been validated to certain degrees in other surgical specialties,3,10,12 they ultimately still rely on the judgment of the examiner. Only a purely quantitative instrument will yield data with minimal inter-rater variability. Additionally, intraoperative information will document sentinel event markers (eg, capsular rupture and vitreous loss) and identify complexities of certain cases. The inclusion of such metrics avoids over-reliance on proxy measures of performance (ie, expert evaluation) and includes some hard outcomes in the evaluation process (ie, patient safety). Clearly, there is inherent value for checklists, global rating scales, and quantitative data in an ideal evaluation tool.

No tool currently available reliably includes each of these elements. However, several tools do exist for some
of these components individually, each with their own limitations. The Eye Surgical Skills Assessment Test (ESSAT) includes a task checklist and global elements. However, the checklist forms do not extend through the entire cataract extraction procedure. Further, it was developed for, and evaluated in, the wet laboratory environment and thus may not be considered reliable to provide information about the resident’s function during unpredictable in vivo procedures.11

The Objective Assessment of Skills in Intraocular Surgery (OASIS) tool7 is primarily a surgical data entry system and was meant to be used in conjunction with the Global Rating Assessment of Skills in Intraocular Surgery (GRASIS) tool, representing objective and subjective criteria, respectively. However, the data collection in the OASIS platform are extensive and not easily portable, thus limiting its utility for higher volume surgical trainees working across multiple institutions.

The GRASIS global surgical competency form has demonstrated good face and construct validity.8 However, important task-specific elements are not included in this tool and its validity has not been studied extensively. Only the Subjective Phacoemulsification Skills Assessment (SPESA)9 and the Objective Structured Assessment of Cataract Surgical Skill (OSACSS)10 tools combine task-specific criteria together with a global index scale. They are both promising tools; however, neither has been assessed for both validity and reliability.9,10 Additionally, these scales do not include quantitative data collection elements.

The SEPOCS scale is intended to fill in some of these gaps. It combines task-specific, global, and quantitative elements into a single validated evaluation tool that is clear and simple to use and takes less than 5 minutes to complete. This investigation demonstrated the psychometric properties of the SEPOCS; it has proven to be highly reliable in its ability to distinguish between a single novice case from a single expert case. Clearly, additional trials with larger groups of trainee and expert surgeons will further ensure confidence for its use across multiple learners at differing skill levels.

The SEPOCS tool can be ideally used in an encounter card system, adding assessment to the end of each surgical day. Such systems have been successfully employed in other areas, including emergency medicine13 and obstetrics.14 There are many advantages of such a system. Although the form can be completed quickly, the act of filling in the evaluation tool will provide a structure and opportunity for immediate feedback, which is known to improve learning.15 It will also serve as a formative evaluation if plotted over time, allowing residents to understand their progress and focus on areas of improvement. Additionally, it can provide data for summative evaluations and allow program administrators to understand the distribution of teaching and learning in the institution. Finally, data from multiple trainees as they pass through the program will eventually serve as the basis for the formation of competency benchmarks in surgical education.

The SEPOCS scale as presented in this article is a reliable evaluation tool for the assessment of cataract surgical skill with good initial face and content validity. We hope this tool will assist educators in maximizing the utility of their trainee assessment protocols.

REFERENCES