Eye (2021) 35:1833–1849
https://doi.org/10.1038/s41433-021-01463-8

The ROYAL COLLEGE of OPHTHALMOLOGISTS

REVIEW ARTICLE

Validity of scoring systems for the assessment of technical and non-technical skills in ophthalmic surgery—a systematic review

Thomas Charles Wood[1] · Sundas Maqsood[1] · Mayank A. Nanavaty[1,2] · Saul Rajak[1,2]

Received: 28 August 2020 / Revised: 12 January 2021 / Accepted: 9 February 2021 / Published online: 1 March 2021

Abstract

Evaluation and recommendation of the scoring systems for technical skills (TS) and non-technical skills (NTS) assessments in ophthalmic surgery. A literature search was performed between December 2019 and May 2020. Studies describing the development or validation of TS or NTS scoring systems in ophthalmic surgery were included. Only scoring systems for completion by hand were included. The primary outcome was the validity and reliability status for each scoring system. The secondary outcome was recommendation based on modified Oxford Centre for Evidence-Based Medicine guidelines. Nineteen and five scoring systems were identified for TS and NTS respectively. TS scoring systems exist for cataract surgery (including the steps of phacoemulsification and paediatric cataract surgery) ptosis, strabismus, lateral tarsal strip, vitrectomy, and intraocular surgery in general. NTS scoring systems apply to cataract surgery or ophthalmic surgery in general. No single scoring system satisfied all validity and reliability measures. The recommended TS scoring systems are 'International Council of Ophthalmology's Ophthalmology Surgical Competency Assessment Rubrics' (ICO-OSCAR) for phacoemulsification, strabismus and paediatric cataract surgery, and 'Objective Structured Assessment of Cataract Surgical Skill' (OSACSS). Non-Technical Skills for Surgeons (NOTSS), Observational Teamwork Assessment for Surgery (OTAS) and Anaesthetists Non-Technical Skills (ANTS) are recommended for NTS. There is a paucity of NTS scoring systems. Further research is required to validate all scoring systems to consistent standards. Limitations of the assessment tools included infrequent quantification of face and content validity, and inconsistency in terminology and statistical methods between studies.

Introduction

The acquisition of surgical skills through the Halstedian model of '*see one, do one, teach one*,' is no longer compatible with modern surgical training [1]. The skills that ophthalmic surgery trainees must develop are intricate and challenging. The open, microsurgical and endoscopic techniques required of trainees frequently have steep learning curves [2, 3]. International training institutions such as the Royal College of Ophthalmologists (RCOphth, United Kingdom) and the Accreditation Council for Graduate Medical Education (ACGME, American College of Surgeons) are increasingly focussed on utilising competency based methods of training and assessment, and not merely numerical targets of completed procedures [4–6].

In addition to obtaining the vitally important technical skills (TS) required for surgical procedures, demonstrating competence in non-technical skills (NTS) is also fundamental for the safe and effective surgeon. TS are the intentional psychomotor actions performed by the surgeon intraoperatively (such as instrument and tissue handling), whilst NTS are the cognitive, social and behavioural capabilities underpinning these technical and procedural elements [7]. NTS deficiencies contribute significantly to surgical error, of which 43% are attributed to communication failures alone [8, 9]. Ophthalmology has been identified as a significant contributor to surgical errors secondary to NTS failures, including wrong intraocular lens implantation and the administration of local anaesthetic to the incorrect eye [10, 11]. NTS are recognised as core competencies by the Royal Australasian College of Surgeons, Royal College of Physicians and Surgeons of Canada, ACGME, and the Royal College of Surgeons (UK),

✉ Thomas Charles Wood
tomwoodresearch@gmail.com

[1] Sussex Eye Hospital, Brighton and Sussex University Hospitals NHS Trust, Brighton, UK

[2] Brighton and Sussex Medical School, Falmer, Brighton, UK

but their presence in ophthalmic surgery research and education remains limited [12–16].

In order for the ophthalmic surgical trainee to demonstrate their TS and NTS competencies according to the requirements of international surgical education bodies, appropriate, valid and reliable assessment tools are required. The objectives of this systematic review were to outline the scoring systems for TS and NTS assessments specific to ophthalmic surgery, present the validity and reliability statuses of each scoring system, and make informed recommendations based on these factors.

Methods

This review was conducted in accordance with the guidelines outlined in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement [17].

Information sources and search terms

A comprehensive search of the English language literature on PubMed, ScienceDirect and Cochrane Library was performed between 20th December 2019 and 27th May 2020. Specific search terms entered into all three databases were; 'Ophthalmology AND non-technical AND assessment', 'Ophthalmology AND non-technical AND assessment', 'Ophthalmology AND scoring system,' 'Ophthalmology AND technical AND assessment', 'Ophthalmology AND NOTSS', 'Ophthalmology AND OTAS', 'Ophthalmology AND NOTECHS', 'Ophthalmology AND ANTS', 'Ophthalmology AND SPLINTS', 'Ophthalmology AND ICO-OSCAR', 'Ophthalmology AND OSACCS', 'Ophthalmology AND OSATS', 'Ophthalmology AND ICSAD'. These terms were chosen to incorporate a wide range of studies, and to elicit the specific scoring systems already known to the authors. No limits were applied for publication dates. Each article selected for full text review underwent a reference review; relevant articles that had not previously been elicited by the search terms were included until study saturation occurred.

Study eligibility criteria

Empirical studies describing the development or validation of a scoring system for TS or NTS in ophthalmic surgery were included. The included scoring systems were required to be printable and able to be completed by hand. This ensured that future assessments could be completed in real-time in either simulated or live settings, addressed the fact that computers in the operating theatre are often in use or unavailable, and allows the observed surgeon to take the form away for learning and reflection. Scoring systems that were entirely computer based were therefore excluded. Letters and editorials were included if they provided detailed explanation of their study's methods and results. Articles were excluded if they encompassed specialties not limited to ophthalmic surgery, tools for the clinical assessment of a patient's vision/anatomy/physiology/pathology, and the exclusive validation of simulation models. Non-English language articles, previous reviews, books, and presentations were excluded.

Study selection and data collection

One reviewer (TCW) performed the searches and data extraction. Whenever studies caused ambiguity, their relevance as per the inclusion criteria was discussed amongst all co-authors in order to reach a final agreement on their inclusion. Abstract review was performed for all studies elicited by the search terms. The full text of each article was obtained and scrutinised if the title or abstract revealed at least one of the following points; scoring system, technical, non-technical, skill, training, assessment, development or validation. Duplicates were removed at this stage. Full texts meeting the inclusion criteria were reviewed for data extraction.

Outcome measures

The primary outcome measure was the validity and reliability status for each scoring system, which was evaluated in accordance with pre-set definitions (Table 1) [18–21].

Table 1 Definitions of validity and reliability for assessment tools [18–21].

Parameter	Definition
Face validity	The extent to which the examination resembles its corresponding real-world situation
Content validity	The extent to which the intended domain is measured by the assessment
Construct validity	The extent to which an assessment is able to differentiate between those of different abilities or experience levels
Concurrent validity	The extent to which the results of the assessment correlate with the gold standard tests known to measure the same domain
Predictive validity	The extent to which the assessment will predict future performance
Interrater reliability/agreement	The extent to which the results obtained by two or more assessors agree for the same participant.
Internal consistency	The extent of item homogeneity within an assessment tool
Educational impact	The extent to which the results and feedback are able to improve the trainee's learning experience

Table 2 Modified Educational Oxford Centre for Evidence-Based Medicine (OCEBM) levels of evidence (A) and levels of recommendation (B) [22].

A

LoE	Criteria
1a	Systematic reviews (meta-analysis) containing at least some trials of level 1b evidence, in which results of separate, independently conducted trials are consistent
1b	Randomised controlled trial of good quality and of adequate sample size (power calculation)
2a	Randomised trials of reasonable quality and/or of inadequate sample size
2b	Nonrandomized trials, comparative research (parallel cohort)
2c	Nonrandomized trial, comparative research (historical cohort, literature controls)
3	Nonrandomized, noncomparative trials, descriptive research
4	Expert opinions, including the opinion of Work Group members

B

LoR	Criteria
1	Based on one systematic review (1a) or at least two independently conducted research projects classified as 1b
2	Based on at least two independently conducted research projects classified as level 2a or 2b, within concordance
3	Based on one independently conducted research project level 2b, or at least two trials of level 3, within concordance
4	Based on one trial at level 3 or multiple expert opinions, including the opinion of Work Group members (e.g. level 4)

The secondary outcome measure included recommendation based on formal criticism in accordance with modified Oxford Centre for Evidence-Based Medicine guidelines. Levels of recommendation (LOR) were provided based on the guideline's levels of evidence (LOE) (Table 2A, B) [22]. The methodology of each study was critiqued in order to reveal strengths and limitations. Risk of bias assessments were conducted for all studies; recognised forms of study bias were stated wherever they were identified. A formal risk of bias assessment tool was not utilised for this review, given the heterogenous nature of studies elicited.

Statistical analysis

Data from all included studies was tabulated. Articles were classified according to their emphasis on TS or NTS. Data extracted from each study included the scoring system used, analysis in a simulated or live setting, participant numbers and their training levels, and the validity and reliability statuses obtained. The heterogenous nature of the development and validation of these scoring systems meant that direct statistical comparisons and meta-analyses were neither appropriate nor applicable.

Results

Study selection

Eight hundred and forty potentially relevant articles were identified through the database searches. Seven hundred and thirty-five abstracts were then reviewed and excluded. One hundred and five articles underwent a full text review, after which 78 irrelevant or duplicate articles were excluded. Therefore, 27 articles merited final inclusion (Fig. 1). From these 27 articles, 19 assessment tools for TS and 5 assessment tools for NTS were identified.

Outline of included TS studies

International Council of Ophthalmology's Ophthalmology Surgical Competency Assessment Rubrics (ICO-OSCAR)

ICO-OSCARs are TS scoring systems, which divide an ophthalmic surgical procedure into task specific and generic components. Objective performance measures use the Dreyfus scale of skill acquisition, whereby numerical scores correlate with a competence level (e.g. Novice = 2, Competent = 5). ICO-OSCARS are freely available to download and have been translated into multiple languages [23].

(a) Extracapsular cataract surgery (ECCE): Presented as a letter to the editor, ICO-OSCAR: ECCE contains fourteen task specific and six global indices of assessment [24]. Twelve international content experts granted face and content validity [24].
(b) Small Incision Cataract Surgery (SICS): The Sim-OSCCAR:SICS was developed from the ICO-OSCAR:SICS template [25]. The study which originally developed ICO-OSCAR:SICS was not found, despite the tool being available on the International Council of Ophthalmology's website [23]. It contains fourteen task specific components and six global indices, and is specific for use in simulated settings. Face and content validity were granted by a panel of twelve international experts using Likert scales (4.6/5 and 4.5/5, respectively) [25]. Four expert surgeons assessed eight cataract

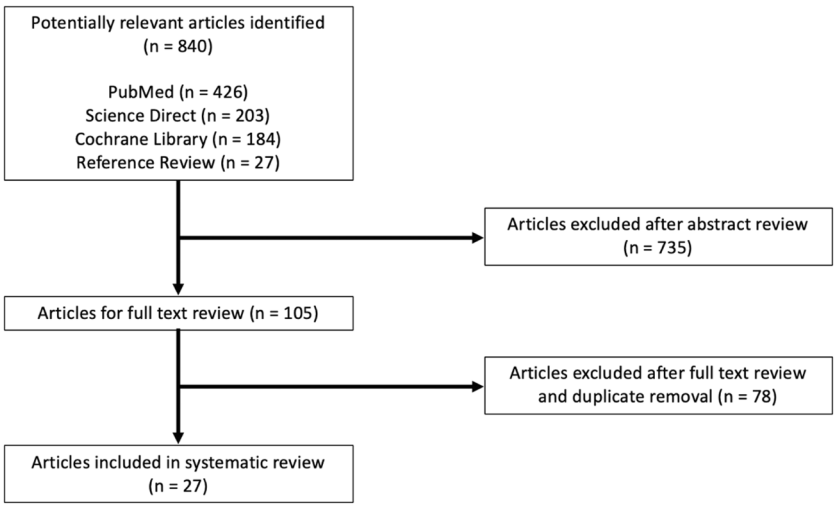

Fig. 1 Flow chart of study selection. Flow chart of article identification, article exclusion, full text review, and article inclusion.

surgeons in a simulated setting. Interrater reliability was assessed using a Krippendorff alpha calculation, with seventeen of the twenty components demonstrating $\alpha > 0.6$ (the level deemed acceptable) [25]. Construct validity was demonstrated as competent surgeons outperformed novices when analysed with a Wilcoxon rank-sum test (novices 0.5–3.25/40, competent 21.5–36.5/40, $p = 0.02$) [25].

(c) Phacoemulsification: ICO-OSCAR:phaco is a twenty component TS scoring system for the assessment of phacoemulsification, which achieved face and content validity according to fifteen international content experts [24, 26]. The assessment of residents of different abilities following six recorded live phacoemulsification procedures demonstrated high internal consistency overall ($\alpha = 0.92$), whilst seventeen components demonstrated an $\alpha > 0.7$ [24, 26].

(d) Paediatric Cataract Surgery: ICO-OSCAR:Paediatric Cataract Surgery is a twenty-two component scoring system for the assessment of paediatric cataract surgery, with fourteen task specific and eight global components; one of which is communication [27]. It could be assumed from the editorial that face and content validity were granted by the international panel of experts consulted, however this was not clear in text and cannot be stated here with confidence [27]. One further study analysed the use of ICO-OSCAR: Paediatric Cataract Surgery in video based recordings of forty-two consultant ophthalmic surgeons and thirty-four ophthalmic surgery fellows. Good interrater agreement was demonstrated using Cohen's kappa for all assessed components, including anterior capsulorhexis (95.72% and 0.84), wound construction (98.36% and 0.83) and intraocular lens (IOL) implantation (96.54% and 0.82) [2]. Construct validity could not be demonstrated due to there being no significant differences in scores obtained between consultants and fellows, however it was recognised that evidencing construct validity at higher levels of training can be difficult [2].

(e) Trabeculectomy: ICO-OSCAR:trabeculectomy is a twenty component TS scoring system for the assessment of trabeculectomy surgery [28]. Thirteen components are specific to the steps of trabeculectomy, whilst seven are global indices (including one knowledge and one communication based component). ICO-OSCAR:trabeculectomy was granted face and content validity by ten international content experts [28].

(f) Vitrectomy: ICO-OSCAR:Vitrectomy is a twenty component scoring system for assessment of the steps of vitrectomy, which was granted face and content validity by eight content experts [29]. This study was presented as an editorial, outlining the development of the tool with well described methodology [29].

(g) Strabismus: ICO-OSCAR:strabismus is a seventeen component TS scoring system for the assessment of strabismus surgery [30]. Eleven components are specific to strabismus procedures, whilst six components are global indices, including one knowledge and one communication element. Its developmental study determined face and content validity through seven content experts using Likert scales [30]. A more recent study of five residents performing strabismus surgery demonstrated the tool's high interrater agreement overall (Cronbach $\alpha = 0.9$), with all but one component achieving Cronbach $\alpha > 0.7$ [31].

(h) Ptosis: OSCAR:ptosis was developed as the OSCAR for the assessment of anterior approach ptosis surgery, with seventeen components of TS assessment [32]. An

international panel of content experts approved its face and content validity [32].

(i) Lateral Tarsal Strip (LTS): ICO-OSCAR:LTS is a seventeen component scoring system for the assessment of lateral tarsal strip surgery, containing nine task specific components and eight global indices [33]. Developed initially by seven content experts, the first draft was reviewed by eleven international content experts who granted face and content validity [33].

Objective structured assessment of technical skills (OSATS) and objective structured assessment of cataract surgical skill (OSACSS)

The video based modified OSATS contains four TS domains of assessment [34, 35]. Scores obtained from a Likert scale of 1–5 correlate with procedure specific requirements [34, 35]. Fourteen resident ophthalmic surgeons underwent a simulated corneal suturing course; the interrater reliability of the modified OSATS was Cronbach $\alpha = 0.78$ [35]. Concurrent validity was demonstrated for the modified OSATS when the motion tracking 'Imperial College Surgical Assessment Device' (ICSAD) was used for the gold standard comparison [35, 36]. Spearman's Rank correlated the combined OSATS scores and ICSAD parameters; OSATS scores correlated significantly with path length ($r = -0.765$, $p < 0.01$), hand movements ($r = -0.55$ $p < 0.01$) and time (time $r = -0.631$, $P < 0.01$) [35].

OSACSS is a twenty component TS scoring system with fourteen task specific and six global components specific to cataract surgery [37]. It is measured on a five point Likert scale [37]. Construct validity was proven for trainee surgeons of lower experience levels in live phacoemulsification surgery; significant differences were demonstrated between the groups with <50 and 50–249 procedures respectively ($p = 0.002$), and between the groups with 50–249 and 250–500 procedures ($p = 0.003$) [37]. Good interrater reliability was demonstrated in a study of nineteen cataract surgeons of varying experience levels, whose recorded phacoemulsification performances were assessed pre and post training (Cronbach $\alpha = 0.92$ and 0.86 respectively), however this version of OSACSS was modified as the draping component was omitted [38].

When used in studies to validate virtual reality simulation modules, OSACSS and the modified OSATS further demonstrated their interrater reliabilities through the intraclass correlation coefficient ($r = 0.788$ for OSACSS capsulorhexis, $r = 0.764$ for OSATS) [39, 40]. Modified OSATS' and OSACSS' construct validities were demonstrated by showing significant differences in scores obtained between cataract surgeons and medical students; OSATS ($p = 0.001$), OSACSS capsulorhexis ($p = 0.003$), hydromaneuvres ($p = 0.017$), phacoemulsification ($p = 0.001$) [40].

Subjective phacoemulsification skills assessment (SPESA)

Presented as a letter to the editor, SPESA is specific to phacoemulsification and contains thirteen TS components, with further components for knowledge, flow and complication management [41]. Assessments are made on a Likert scale of 1–5. Interrater reliability was reported without alpha calculations, with 85% of 9/12 of the components falling within one standard deviation of the mean [41].

Ophthalmic plastic surgical skills assessment tool (OPSSAT)

OPSSAT is a TS assessment tool with eighteen components specific for ophthalmic plastic surgery, with one communication component [42]. Scores are generated on a Likert scale of 1–5. It was granted face and content validity by twenty ophthalmic plastic surgeons, with 90% in agreement regarding its component weighting and content [42].

Strabismus surgical skills assessment tool

The strabismus surgical skills assessment tool contains seventeen components of assessment, with one knowledge and one communication based element [43]. Assessments are made on a Likert scale of 1–5. Face and content validity were granted by twenty strabismus surgeons who refined the tool [43].

Global rating assessment of skills in intraocular surgery (GRASIS)

GRASIS contains eleven components relevant to intraocular surgery, four of which are non-technical [44]. Assessments are made on a Likert scale of 1–5. Face and content validity were granted by twenty-two educational experts, with all components of assessment achieving at least a 'very useful' rating [44].

Ophthalmology wet lab structured assessment of skill and technique scoring rubric (OWLSAT)

The University of Iowa Department of Ophthalmology Wet Laboratory Structured Assessment of Skill and Technique Scoring Rubric (OWLSAT) was developed by content experts and validated by an external task force [45]. OWLSAT was utilised in another simulation study, however the validity and reliability were not explored [46]. The aims of these studies was to produce a simulation programme, however the development and validity outcomes of these scoring systems were minimally described [45, 46].

Surgical skills assessment rubric for pterygium surgery

The Surgical skills assessment rubric for pterygium surgery is a sixteen-point system specific to the steps of pterygium surgery, with residents scored according to the Dreyfus scale. Face and content validities were achieved. 2 blinded assessors assessed 12 residents during live surgery in order to establish the tool's interrater reliability; the intraclass correlation coefficient was 0.90 (95% CI, 0.76–0.96, $p < 0.001$). Furthermore, resident scores later in rotation were significantly higher than those obtained earlier (4.32 ± 0.35 vs. 9.36 ± 0.31, $p = 0.006$), however formal construct validity testing did not occur [47].

Further evaluation tools

The 'Evaluation tool for Smith et al. Evaluation of Capsulorhexis Technique' contains a mixture of fourteen TS components from GRASIS and ICO-OSCAR:phaco, with a range of assessment options depending on where the components originated [26, 44, 48]. Each component's reliability and validity was assessed individually with no analysis of the tool overall, however the intraclass correlation coefficients (ICC) for components such as 'flow of operation' (ICC 0.87, $p < 0.028$) and 'commencement of flap' (ICC 0.78, $p < 0.004$) were high for interrater reliability [48].

The 'Evaluation Form for Smith et al. Surgical Technique' contains fifteen TS components from GRASIS and ICO-OSCAR:phaco, for hydrodissection and phacoemulsification [26, 44, 49]. Each component's reliability and validity was assessed individually with no analysis of the tool overall, however the ICC (interrater reliability) was high for components such as 'instrument handling during hydrodissection' (ICC 0.71, $p < 0.0001$) and 'flow of operation: time and motion during hydrodissection' (ICC 0.72, $p < 0.0001$) [49].

Outline of NTS studies

Saleh et al. analysed the use of NTS assessment tools within ophthalmic surgery, all of which were developed and validated without being specific to any particular speciality [50]. These were Observational Teamwork Assessment for Surgery (OTAS), Non-Technical Skills Scale (NOTECHS), Non-Technical Skills for Surgeons (NOTSS) and Anaesthetists Non-Technical Skills (ANTS) [50–54]. OTAS assesses teamwork behaviours, with components including communication, coordination, cooperation, leadership and situational awareness across perioperative and intraoperative periods [51]. NOTECHS assesses communication and interaction, situational awareness and vigilance, cooperation and team skills, leadership and managerial skills, and decision making [54]. NOTSS assesses the surgeon's situational awareness, decision making, task management, leadership, communication and teamwork [52]. ANTS assesses the task management, teamwork, situational awareness and decision making of anaesthetists [53].

Twenty simulations of surgical teams managing complicated scenarios were used to cross validate the tools using the Pearson product moment correlation coefficient. Normalised standard deviations demonstrated interrater reliability for NOTSS (0.024, 95% CI, 0.014–0.091), ANTS (0.068, 95% CI, 0.041–0.194), OTAS (0.060, 95% CI, 0.034–0.225) and NOTECHS (0.072, 95% CI, 0.043–0.206) [50]. Concurrent validity was obtained through correlating scores obtained with each tool. ANTS, NOTSS and OTAS achieved content validity and internal consistency, however NOTECHS was deemed the least applicable [50].

The HUman Factors in intraoperative Ophthalmic Emergencies Scoring System (HUFOES) was developed using Delphi methodology. Content validity was granted by 14 ophthalmic surgeons, with 85.7% ($n = 12$) respondents in agreement that HUFOES components can accurately identify and assess the listed NTS [55]. Furthermore, HUFOES' construct validity has been proposed, with 78.6% ($n = 11$) of respondents in agreement that HUFOES has the ability to distinguish between those of different training levels [55]. HUFOES was developed and validated as the first NTS scoring system for managing intraoperative emergencies in cataract surgery, using posterior capsular rupture as an example [55]. The authors have stated that further research is required for rigorous assessment of HUFOES' interrater reliability, internal consistency, construct and concurrent validities [55].

Evaluation of the included studies

Primary outcome measures

The primary outcomes extracted from each study are presented in Table 3 Study Outcomes for TS Scoring Systems, and Table 4 Study Outcomes for NTS Scoring Systems. These tables include the skillset, the intended subspecialty, study type or setting, the scoring system evaluated, and the primary outcome for each study. An overview of the validity and reliability status obtained by each scoring system is displayed in Table 5.

Secondary outcome measures

Secondary outcome measures of recommendation relating to the strengths, limitations, OCEBM status and risk of bias assessments for all included studies are presented in Table 6.

Table 3 Study outcomes for TS scoring systems.

Study/[Ref.]	Skillset	Subspecialty	Study/setting	Participants	Scoring system	Primary outcomes
Golnik et al. [24]	TS	Cataract surgery	Developmental	12 Content experts	ICO-OSCAR:phaco	Face and content validity granted but not quantified.
	TS				ICO-OSCAR:ECCE	Face and content validity granted but not quantified.
Dean et al. [25]	TS	Small incision cataract surgery	Simulated		Sim-OSSCAR:SICS	Face and content validity granted with Likert scales (4.6/5 and 4.5/5 respectively). Interrater reliability demonstrated with Krippendorff alpha; 17/20 components demonstrated $\alpha > 0.6$. Construct validity demonstrated with Wilcoxon rank-sum test; competent surgeons outperformed novices (novices 0.5–3.25/40, competent 21.5–36.5/40, $p = 0.02$).
Golnik et al. [26]	TS	Phacoemulsification surgery	Live	15 Content experts	ICO-OSCAR:phaco	Face and content validity agreed by consensus. High internal consistency score ($\alpha = 0.92$). 17 components achieved $\alpha > 0.7$.
Badakere et al. [2]	TS	Paediatric cataract surgery	Live	42 Consultant videos 34 fellow videos	ICO-OSCAR:Paediatric cataract surgery	Interrater agreement demonstrated using Cohen's kappa for all components; anterior capsulorhexis (95.72% and 0.84), wound construction (98.36% and 0.83) and intraocular lens (IOL) implantation (96.54% and 0.82). Construct validity not proven.
Swaminathan et al. [27]	TS	Paediatric cataract surgery	Developmental	International expert panel	ICO-OSCAR:Paediatric cataract surgery	Possible to assume face and content validity but neither specified nor quantified.
Green et al. [28]	TS	Trabeculectomy surgery	Developmental	10 Content experts	ICO-OSCAR:Trabeculectomy	Face and content validity granted but not quantified.
Golnik et al. [29]	TS	Vitrectomy surgery	Developmental	8 Content experts	ICO-OSCAR:vit	Face and content validity granted but not quantified.
Golnik et al. [30]	TS	Strabismus surgery	Developmental	7 Content experts	ICO-OSCAR:strabismus	Face and content validity granted but not quantified.
Motley et al. [31]	TS	Strabismus surgery	Live	5 Resident surgeons 10 expert assessors	ICO-OSCAR:strabismus	High interrater agreement overall (Cronbach $\alpha = 0.9$). All but one component achieved Cronbach $\alpha > 0.7$.
Juniat et al. [32]	TS	Anterior approach ptosis surgery	Developmental	International expert panel	OSCAR:ptosis	Face and content validity agreed by consensus.
Golnik et al. [33]	TS	Lateral tarsal strip surgery	Developmental	11 Content experts	ICO-OSCAR:LTS	Face and content validity granted but not quantified.
Selvander and Asman [39]	TS	Cataract surgery	Simulated	17 Medical students. 7 cataract surgeons.	OSACCS	OSACSS and modified OSATS demonstrated interrater reliabilities through the intraclass correlation coefficient ($r = 0.788$ for OSACSS capsulorhexis, $r = 0.764$ for OSATS).
	TS				OSATS	

Table 3 (continued)

Study/[Ref.]	Skillset	Subspecialty	Study/setting	Participants	Scoring system	Primary outcomes
Selvander and Asman [40]	TS TS	Hydromaneuvres and phacoemulsification	Simulated	17 Medical students. 7 cataract surgeons.	OSACCS OSATS	Construct validities demonstrated; significant differences in scores obtained between cataract surgeons and medical students; OSATS ($p = 0.001$), OSACSS capsulorhexis ($p = 0.003$), hydromaneuvres ($p = 0.017$), phacoemulsification ($p = 0.001$).
Saleh et al. [37]	TS	Phacoemulsification surgery	Live	38 Ophthalmic surgeons with varying experience. 3 Blinded assessors.	OSACSS	Construct validity demonstrated for surgeons of lower experience levels. Significant differences demonstrated between groups with <50 and 50–249 procedures ($p = 0.002$), and between groups with 50–249 and 250–500 procedures ($p = 0.003$).
Thomsen et al. [38]	TS	Phacoemulsification surgery	Simulated	19 Cataract surgeons. 3 blinded assessors.	OSACSS	Interrater reliability demonstrated using Cronbach alpha. Pre and post training scores; Cronbach $\alpha = 0.92$ and 0.86 respectively.
Ezra et al. [35]	TS	Corneal suturing	Simulated	14 Residents	Modified OSATS	Interrater reliability; Cronbach $\alpha = 0.78$. Concurrent validity demonstrated with Spearman's Rank correlating combined OSATS scores and ICSAD parameters; OSATS scores correlated significantly with path length ($r = -0.765$, $p < 0.01$), hand movements ($r = -0.55$ $p < 0.01$) and time (time $r = -0.631$, $P < 0.01$).
Feldman and Geist [41]	TS	Phacoemulsification	Live	4 Attending and 1 resident surgeon. 14 Assessors.	SPESA	Interrater reliability was reported without alpha calculations, with 85% of 9/12 of the components falling within one standard deviation of the mean.
Gauba et al. [42]	TS	Ophthalmic plastic surgery	Developmental	20 Ophthalmic plastic surgeons	OPSSAT	Face and content validity granted. 90% agreement regarding its component weighting and content.
Pilling et al. [43]	TS	Strabismus surgery	Developmental	20 Expert surgeons	Strabismus surgical skills assessment tool	Face and content validity granted.
Cremers et al. [44]	TS	Intraocular surgery	Developmental	22 Educational experts	GRASIS	Face and content validity granted, with all components of assessment achieving at least a 'very useful' rating.
Gertsch et al. [46]	TS	Strabismus surgery	Simulated	2 Content experts	OWLSAT	No validity or reliability outcomes described.
Lee et al. [45]	TS	Cataract surgery	Developmental	Content experts	OWLSAT	No validity or reliability outcomes described.
Zarei-Ghanavati et al. [47]	TS	Pterygium surgery	Developmental Live	Designed and refined by experienced panel of surgeons. 2 Blinded assessors for 12 residents.	Surgical skills assessment rubric for pterygium surgery	Face and content validity granted but not quantified. Intraclass correlation coefficient 0.90 (95% CI, 0.76–0.96, $p < 0.001$). Resident scores later in rotation were significantly higher than earlier scores (4.32 ± 0.35 vs. 9.36 ± 0.31, $p = 0.006$).

Table 3 (continued)

Study/[Ref.]	Skillset	Subspecialty	Study/setting	Participants	Scoring system	Primary outcomes
Smith et al. [49]	TS	Hydrodissection and phacoemulsification	Live	Expert panel. 5 Surgeons operating with varying experience.	Evaluation tool for Smith et al. Evaluation of capsulorhexis technique	Each component's reliability and validity was assessed individually with no analysis of the tool overall, however the ICC (interrater reliability) was high for components such as 'instrument handling during hydrodissection' (ICC 0.71, $p < 0.0001$) and 'flow of operation: time and motion during hydrodissection' (ICC 0.72, $p < 0.0001$).
Smith et al. [48]	TS	Capsulorhexis surgery	Live	Expert panel. 6 Surgeons operating with varying experience.	Evaluation form for Smith et al. surgical technique	Each component's reliability and validity assessed individually with no overall analysis. Intraclass correlation coefficients (ICC) for components such as 'flow of operation' (ICC 0.87, $p < 0.028$) and 'commencement of flap' (ICC 0.78, $p < 0.004$) were high for interrater reliability.

Scoring system recommendations

ICO-OSCAR:phaco, ICO-OSCAR:strabismus and ICO-OSCAR:Paediatric Cataract Surgery demonstrate LoE 3 and 4, however all ICO-OSCARs currently stand at LOR 4. OSACSS has featured in multiple LOE 3 studies and is therefore recommended at LOR 3 for cataract surgery [37–40].

ANTS, NOTSS and OTAS are NTS assessment tools valid for use in ophthalmic surgery [50–53]. As one LOE 3 study evaluated their use in ophthalmic surgery, they are recommended at LOR 4.

Discussion

In this comprehensive review, the TS and NTS scoring systems specific to ophthalmic surgery were evaluated. Nineteen TS and five NTS scoring systems for assessment in ophthalmic surgery were identified in twenty-seven studies. TS scoring systems exist for cataract surgery (including the specific steps of phacoemulsification, capsulorhexis, and paediatric cataract surgery) ptosis, strabismus, lateral tarsal strip, vitrectomy, and intraocular surgery in general [2, 24–33, 37, 41–44, 48]. The recommended scoring systems for TS are ICO-OSCAR:phaco, ICO-OSCAR:strabismus and ICO-OSCAR:Paediatric Cataract Surgery and OSACSS [37–40]. The scoring systems identified for NTS assessment in ophthalmic surgery are NOTSS, OTAS, ANTS, NOTECHS and HUFOES, of which NOTSS, OTAS and ANTS can currently be recommended [50–53, 55].

The aims of TS scoring systems are to assess surgical skills, enhance learning curve progression, identify strengths and weaknesses, ensure training objectives are met, promote reflective practice and create feedback opportunities [56]. However, scoring systems can be perceived as complex, time consuming, user dependent tick-box exercises [56, 57]. The *Halo effect* can also apply, whereby positive performance early in the assessment promotes a cognitive bias, resulting in the assessor overlooking less favourable performances later on [58]. Scoring systems must therefore demonstrate strong validity and reliability statuses, in order to mitigate against these limitations. For instance, a high interrater reliability provides an assurance against user dependence, subjectivity, and the Halo effect.

The acquisition of surgical skills is particularly challenging in an era where training hours have been substantially reduced [57, 59]. Virtual reality systems providing automated assessments are expanding but remain costly, and therefore direct evaluation of the trainee remains fundamental for the facilitation of skill growth and individualised

Table 4 Study outcomes for NTS scoring systems.

Study/ [Ref.]	Skillset	Subspecialty	Study/setting	Participants	Scoring system	Outcome
Saleh et al. [50]	NTS	Ophthalmic surgery	Simulated	Consultant and trainee ophthalmologists, anaesthetists, operating department practitioners, nurses.	ANTS	Content validity granted for ophthalmic surgery. Concurrent validity achieved by correlating scores obtained with each tool. Interrater reliability using standardised normal deviation (0.068, 95% CI, 0.041–0.194) internal consistency.
					NOTSS	Content validity granted for ophthalmic surgery. Concurrent validity achieved by correlating scores obtained with each tool. Interrater reliability using standardised normal deviation (0.024, 95% CI, 0.014–0.091), internal consistency.
					NOTECHS	Concurrent validity achieved by correlating scores obtained with each tool. Interrater reliability using standardised normal deviation (0.072, 95% CI, 0.043–0.206)
					OTAS	Content validity granted for ophthalmic surgery. Concurrent validity achieved by correlating scores obtained with each tool. Interrater reliability using standardised normal deviation (0.060, 95% CI, 0.034–0.225) internal consistency.
Wood et al. [55]	NTS	Cataract Surgery	Developmental	Proposed by focus group of 2 Consultant ophthalmologists, 1 ophthalmology registrar, and 1 academic doctor. Further developed and validated by 14 consultant ophthalmologists.	HUFOES	All HUFOES components achieved importance rating >8/10 with interrater agreement for the importance of components achieving $\alpha = 0.953$. Content validity granted; 85.7% ($n = 12$) agreed that HUFOES components can identify and assess the listed NTS. Construct validity proposed but not confirmed; 78.6% ($n = 11$) agreed that HUFOES has the ability to distinguish between those of different training levels.

learning [44, 60]. Several European countries are actively refining their surgical training programmes to be competency based, which requires valid and reliable methods through which the trainee's progress can be accurately assessed [5, 57]. To date, studies assessing the skills of ophthalmic surgery trainees have used heterogenous scoring systems without evidence-based recommendations. This is problematic as the strongest tools are those which have had the greatest number of validity and reliability statuses favourably assessed (Table 1). The present review found the ICO-OSCARs to be a robust collective of scoring systems, providing a consistent rating scale and common format across a range of subspecialties. OSCAR:ptosis, ICO-OSCAR:phaco, ICO-OSCAR:strabismus, ICO-OSCAR: trabeculectomy, ICO-OSCAR:LTS, ICO-OSCAR:Vitrectomy, ICO-OSCAR:ECCE, and Sim-OSSCAR:SICS all demonstrated face and content validity, but only Sim-OSSCAR:SICS demonstrated construct validity [24–26, 28–33]. ICO-OSCAR:strabismus, ICO-OSCAR:Paediatric cataract surgery and Sim-OSSCAR:SICS demonstrated interrater reliability, whilst ICO-OSCAR:phaco demonstrated internal consistency [2, 25, 26, 31]. Based on these factors, all of the stated ICO-OSCAR scoring systems are valuable within their intended domains. However, further research is required to comprehensively evaluate the individual validity and reliability statuses that each tool currently lacks.

NTS include teamwork, communication, leadership, situational awareness and stress response, which are fundamental for safe surgery and complication management. NTS failures in ophthalmic surgery are commonly responsible for adverse events including wrong intraocular lens implantation and administration of local anaesthetic to the incorrect eye [10, 11]. Despite this, only two studies have explored and evaluated the use of NTS scoring systems in ophthalmic surgery, one of which focussed on the surgical team as opposed to the surgeon specifically [50]. ANTS, NOTSS and OTAS were developed and validated for other surgical specialties and are therefore not specific to the NTS requirements of ophthalmic surgery. Furthermore, ANTS was originally designed to encompass the NTS required of anaesthetists, but was considered useful by Saleh et al. given that the listed NTS were also applicable to the surgeon [50, 53]. ANTS, NOTSS and OTAS were all found to demonstrate content validity, concurrent validity, internal consistency and interrater reliability in the ophthalmic surgery domain when analysed during simulated recreations of genuine patient safety incidents [50–53]. Whilst they are not tailored to the specific NTS requirements of ophthalmic surgery procedures, it is encouraging that their application was found to be valid in this setting [50, 61]. Furthermore, scoring systems specific to the NTS requirements of ophthalmic surgery are being developed and validated. The recently developed HUFOES has provided a content validated NTS scoring system for

Table 5 Summary of scoring systems and their validity and reliability statuses.

Study/[Ref.]	Skillset	Tool	Face validity	Content validity	Construct validity	Concurrent validity	Predictive validity	Interrater reliability	Internal consistency	Educational impact
Golnik et al. [24]	TS	ICO-OSCAR:phaco	Y	Y	N	N	N	N	N	N
		ICO-OSCAR:ECCE	Y	Y	N	N	N	N	N	N
Dean et al. [25]	TS	Sim-OSSCAR:SICS	Y	Y	Y	N	N	Y	N	N
Golnik et al. [26]	TS	ICO-OSCAR:phaco	Y	Y	N	N	N	Y	N	N
Badakere et al. [2]	TS	ICO-OSCAR:Paediatric Cataract Surgery	N	N	N	N	N	Y	N	N
Swaminathan et al. [27]	TS	ICO-OSCAR:Paediatric Cataract Surgery	N	N	N	N	N	N	N	N
Green et al. [28]	TS	ICO-OSCAR:Trabeculectomy	Y	Y	N	N	N	N	N	N
Golnik et al. [29]	TS	ICO-OSCAR:vit	Y	Y	N	N	N	N	N	N
Golnik et al. [30]	TS	ICO-OSCAR:strabismus	Y	Y	N	N	N	N	N	N
Motley et al. [31]	TS	ICO-OSCAR:strabismus	N	N	N	N	N	Y	N	N
Juniat et al. [32]	TS	OSCAR:ptosis	Y	Y	N	N	N	N	N	N
Golnik et al. [33]	TS	ICO-OSCAR:LTS	Y	Y	N	N	N	N	N	N
Selvander and Asman [39]	TS	OSACSS	N	N	N	N	N	Y	N	N
	TS	OSATS	N	N	Y	N	N	Y	N	N
Selvander and Asman [40]	TS	OSACSS	N	N	Y	N	N	Y	N	N
	TS	OSATS	N	N	Y	N	N	Y	N	N
Saleh et al. [37]	TS	OSACSS	N	N	Y	N	N	N	N	N
Thomsen et al. [38]	TS	OSACSS	N	N	N	N	N	Y	N	N
Ezra et al. [35]	TS	Modified OSATS	N	N	N	Y	N	Y	N	N
Feldman and Geist [41]	TS	SPESA	N	N	N	N	N	Y	N	N
Gauba et al. [42]	TS	OPSSAT	Y	Y	N	N	N	N	N	N
Pilling et al. [43]	TS	Strabismus surgical skills assessment tool	Y	Y	N	N	N	N	N	N
Cremers et al. [44]	TS	GRASIS	Y	Y	N	N	N	N	N	N
Gertsch et al. [46]	TS	OWLSAT	N	N	N	N	N	N	N	N
Lee et al. [45]	TS	OWLSAT	N	N	N	N	N	N	N	N
Zarei-Ghanavati et al. [47]	TS	Surgical skills assessment rubric for pterygium surgery	Y	Y	N	N	N	Y	N	N
Smith et al. [49]	TS	Evaluation form for Smith et al. surgical technique	N	N	N	N	N	N	N	N

Table 5 (continued)

Study/[Ref.]	Skillset	Tool	Face validity	Content validity	Construct validity	Concurrent validity	Predictive validity	Interrater reliability	Internal consistency	Educational impact
Smith et al. [48]	TS	Evaluation tool for Smith et al. Evaluation of capsulorhexis technique	N	N	N	N	N	N	N	N
Saleh et al. [50]	NTS	ANTS	N	Y	N	Y	N	Y	Y	N
		NOTSS	N	Y	N	Y	N	Y	Y	N
		NOTECHS	N	N	N	Y	N	Y	N	N
		OTAS	N	Y	N	Y	N	Y	Y	N
Wood et al. [55]	NTS	HUFOES	N	Y	N	N	N	N	N	N

managing intraoperative emergencies during cataract surgery [55]. HUFOES has demonstrated content validity with preliminary indications of its construct validity, however further analysis is pending [55]. Together, these tools should be used as a basis for the development of further NTS scoring systems in ophthalmic surgery.

OSCAR:ptosis, ICO-OSCAR:phaco, ICO-OSCAR:ECCE, ICO-OSCAR:strabismus, ICO-OSCAR:trabeculectomy, ICO-OSCAR:LTS, ICO-OSCAR:Vitrectomy, ICO-OSCAR:Paediatric cataract surgery, OPSSAT, GRASIS and Strabismus Surgical Skills Assessment Tool demonstrated individual validity and reliability statuses which qualified their relevance within their intended domains, but they were not evaluated further in simulated or live settings [24, 27–30, 32, 33, 42–44]. This was not considered to be detrimental to the outcomes of this review, as studies focussing entirely on the development of a scoring system are not required to analyse them in simulated or live settings. However, evaluation of the scoring systems in simulated or live settings must be undertaken before they can be deemed robust, valid and reliable.

Despite being present occasionally, bias was not found to be widespread in the studies overall, therefore having negligible impact on overall outcomes. OWLSAT, the 'Evaluation tool for Smith et al. Evaluation of Capsulorhexis Technique', and the 'Evaluation Form for Smith et al. Surgical Technique' did not provide data on any form of validity or reliability [45, 48, 49]. OWLSAT was developed as a novel assessment tool to facilitate simulation projects with minimal details provided for its development or validity evaluation, therefore indicating a design bias [45, 46]. Furthermore, design bias was found for the Evaluation Form for Smith et al. Surgical Technique' and the 'Evaluation tool for Smith et al. Evaluation of Capsulorhexis Technique' [48, 49]. These used components of assessment from previously validated scoring systems, and explored the validity and reliability of each component of assessment individually, as opposed to the tool overall [48, 49]. Given the lack of validity and reliability data of each of these tools, together with their bias assessments, they cannot be recommended currently. Badakere et al. recognised that assessment of the videos produced for the evaluation of ICO-OSCAR:Paediatric Cataract Surgery was challenging, given that the assessor was not able to see the complexity of the procedure being performed. This raises the possibility of an assessor bias [2]. Dean et al. raised the possibility of response bias when developing the Sim-OSSCAR:SICS due to their use of open ended responses, however this was largely unavoidable and remains a useful means of gaining unrestricted feedback [25].

Most studies evaluated by the present review were transparent about their limitations, and the need for further research to be undertaken into the scoring system they

Table 6 Study strengths, limitations, OCEBM status and risk of bias assessment.

Study/[Ref.]	Strengths	Limitations	Risk of bias	OCEBM
Saleh et al. [50]	Used real patient safety incidents for each scenario. Appropriate statistical methods.	Assessed the surgical team, not the surgeon specifically. ANTS used to assess surgeons despite being designed for anaesthetists.	Not identified	Level 3
Wood et al. [55]	Developed by focus group and refined by content experts using Delphi methodology. Appropriate statistical methods. Acknowledges need for further validity and reliability analysis.	Not yet analysed in live or simulated settings.	Study conducted by authors of present review	Level 4
Golnik et al. [24]	Developed and refined by international content expert panel.	Not yet analysed in live or simulated settings. Face and content validity achieved but not quantified.	Not identified	Level 4
Dean et al. [25]	Refined by panel of international content experts. Appropriate statistical methods used. Comprehensively addressed study limitations. Face and content validity quantified. Video recordings assessed by blinded assessors.	Recognised that open ended feedback comments are a potential source of response bias.	Potential for response bias. Unavoidable.	Level 3
Golnik et al. [26]	Developed, refined and validated by international panel of content experts.	Face and content validity achieved but not quantified. Residents of different levels assessed but construct validity not assessed.	Not identified	Level 3
Badakere et al. [2]	Blinded assessors used to rate videos. Appropriate statistical methods used.	Cut off not given for level of interrater agreement considered acceptable. Recognised difficulty generating construct validity at higher experience levels. Recognised that assessment of their surgical videos is difficult as the assessor will not know the complexity of the surgery involved.	Possible assessor bias	Level 3
Swaminathan et al. [27]	Developed and refined by international content expert panel.	Acknowledged that paediatric cataract surgery can be performed in many ways. No definite reference to face and content validity. Not assessed in live or simulated settings.	Not identified	Level 4
Green et al. [28]	Developed and refined by international content expert panel. Acknowledges need for further validity analysis.	Face and content validity achieved but not quantified. Not assessed in live or simulated settings.	Not identified	Level 4
Golnik et al. [29]	Developed and refined by international content expert panel.	Not yet analysed in live or simulated settings. Face and content validity achieved but not quantified.	Not identified	Level 4
Golnik et al. [30]	Developed and validated by international content expert panel.	Not yet analysed in live or simulated settings. Face and content validity achieved but not quantified.	Not identified	Level 4
Motley et al. [31]	Appropriate statistical methods used	Small sample size. Does not state if assessors were blinded.	Not identified	Level 3
Juniat et al. [32]	Developed and refined by international content expert panel. Acknowledges need for use and educational impact analysis.	Not yet analysed in live or simulated settings. Face and content validity achieved but not quantified.	Not identified	Level 4
Golnik et al. [33]	Developed and validated by international content expert panel.	Validity and reliability not quantified. Not analysed in live or simulated settings.	Not identified	Level 4
Selvander and Asman [39]	Video recordings randomised and assessed by blinded assessors.	Large differences in participant skillsets (experienced surgeons and medical students) limited finer detail. OSATS and OSACCS data difficult to distinguish from each other.	Not identified	Level 3
Selvander and Asman [40]	Video recordings assessed by blinded assessors. Appropriate statistical methods used.	Large differences in participant skillsets (experienced surgeons and medical students) limited finer detail.	Not identified	Level 3
Saleh et al. [37]	Video recordings assessed by blinded assessors.	No definitive reference to face and content validity attainment. Potential to assess further forms of validity but this was not undertaken.	Not identified	Level 3
Thomsen et al. [38]	Masked raters and outcome assessors assessed randomised videos. Appropriate statistical methods used.	No formal analysis into construct validity. No control group (judged to be unethical to include one)	Not identified	Level 3
Ezra et al. [35]	Video recordings assessed by blinded assessors. Appropriate statistical methods used.	ICSAD presumed as gold standard for comparative purposes, however it does have good construct validity for corneal suturing.	Not identified	Level 3
Feldman and Geist [41]	Developed by content experts. Videos assessed by blinded assessors. Acknowledges need for further validity and reliability analysis.	Interrater reliability not assessed using alpha calculation; means and standard deviations used instead.	Not identified	Level 4

Table 6 (continued)

Study/[Ref.]	Strengths	Limitations	Risk of bias	OCEBM
Gauba et al. [42]	Developed by panel of content experts. Face and content validity quantified.	Not analysed in live or simulated settings.	Not identified	Level 4
Pilling et al. [43]	Validated by content expert panel. Extent of agreements quantified.	Not analysed in live or simulated settings.	Not identified	Level 4
Cremers et al. [44]	Developed according to content experts.	Not analysed in live or simulated settings.	Not identified	Level 4
Gertsch et al. [46]	Appropriate statistical methods used.	Scoring system produced without validation to supplement simulation programme, despite acknowledging scoring systems that may have been useful for assessments.	Design bias	Level 3
Lee et al. [45]	Assessment tool and global evaluation form developed by content experts and externally validated by task force.	Scoring system produced to supplement simulation programme with minimal description regarding its development. No validity outcomes presented.	Not identified	Level 4
Zarei-Ghanavati et al. [47]	Developed by content expert panel. Appropriate statistical methods. Assessed interrater reliability in videos of live pterygium surgery	Face and content validity not quantified.	Not identified	Level 3
Smith et al. [49]	Developed by content expert panel. Blinded assessors used to rate videos.	Components of assessment used from previously validated tools. Validity terms not consistent with those used in wider literature. Validity explored for each individual question with no overall analysis of the scoring system itself.	Design bias	Level 3
Smith et al. [48]	Developed by content expert panel. Blinded assessors used to rate videos.	Components of assessment used from previously validated tools. Validity terms not consistent with those used in wider literature. Validity explored for each individual question with no overall analysis of the scoring system itself.	Design bias	Level 3

developed or validated. One common limitation applicable to many of the OSCAR studies was that expert panels frequently granted face and content validity without quantifying the extent to which this was the case [24, 26, 28–30, 32, 33, 47]. Quantification with Likert scales would have been preferable, however this was not considered to have a detrimental impact on the findings of this review given that it is acceptable to achieve face and content validity through expert opinion alone.

This review has limitations. Meta-analysis of the data was not possible due to heterogeneity of study types and outcome metrics. There were differences in the statistical methods used between studies which were often complex and unclear, including when they were assessing the same form of validity or reliability. Furthermore, there is a paucity of published randomised controlled trials and level 1 evidence, thus reducing the level at which the assessment tools can be recommended. Therefore, unlike other quantitative reviews, letters and editorials were included in this review on the condition that the methods and results of the study described were explained comprehensively.

Nonetheless, this review has identified key points for future research, and areas to strengthen the development and validation methods for future TS and NTS training tools in ophthalmic surgery. Firstly, the statistical methods and terminologies by which an assessment tool's validity and reliability are analysed should be standardised, in order to facilitate easier comparisons between studies. Terms and statistics for assessing identical outcomes should be standardised across studies, whilst cut-offs for acceptable levels of interrater reliability should be stated wherever possible. Furthermore, face and content validities should be more frequently quantified. Newly developed scoring systems should be analysed in live or simulated settings to further assess their validity and reliability statuses. Studies focussing on simulation programmes should use previously validated scoring systems for participant assessments, rather than to create novel and lesser valid tools within the same study. No data was provided for the educational impact of the scoring systems identified in this review, which should be a focus for future research. Finally, further NTS assessment tools specific to ophthalmic surgery should be developed and validated, given the expanding recognition of the importance of NTS within this specialty [16].

Taking the primary and secondary outcomes of this review into account, recommendations can be made for TS and NTS scoring systems specific to ophthalmic surgery. ICO-OSCARs are comprehensive scoring systems for TS, however further research needs to be undertaken to analyse them beyond face and content validity alone. ICO-OSCAR: phaco, ICO-OSCAR:strabismus and ICO-OSCAR:Paediatric Cataract Surgery demonstrate LoE 3 and 4, however all ICO-OSCARs currently stand at LOR 4. OSACSS has

featured in multiple LOE 3 studies and is therefore recommended at LOR 3 for cataract surgery [37–40]. ANTS, NOTSS and OTAS are currently the only NTS scoring systems which can be recommended for ophthalmic surgery [50–53]. They were not originally developed for ophthalmic surgery, and so far only one LOE 3 study has evaluated their use in this setting. However, they have been extensively validated elsewhere, and it is likely that their full potential in ophthalmic surgery is yet to be recognised [61]. Given its recent development in a single LoE 4 study, HUFOES has not yet achieved an OCEBM LoR score. Unlike ANTS, NOTSS and OTAS, HUFOES was specifically designed for ophthalmic surgery and has already obtained content validity [55]. Future research will elicit further aspects of HUFOES' potential, with emphasis on its validities, reliability and educational impact.

Conclusion

Scoring systems for TS and NTS have been developed and validated for use in ophthalmic surgery, however their validity and reliability statuses have been evaluated to different extents. Tools exist to satisfy training requirements for multiple domains of ophthalmic surgery, however further research is required to validate them all to consistent standards. This review underlines the need for further research into NTS for ophthalmic surgery and recommends that specific NTS scoring systems are developed and validated for this domain.

Summary

What was known before

- There is an increasing focus on competency based methods of surgical training and assessment.
- The safe and effective surgeon must demonstrate NTS in addition to TS.
- Appropriate, valid and reliable scoring systems are required for accurate surgical skills assessments.

What this review adds

- Nineteen scoring systems for TS assessment and five scoring systems for NTS assessment were identified.
- No single scoring system satisfies all measures of validity and reliability.
- There is a paucity of scoring systems for NTS when compared to those for technical skills.

Author contributions TCW recognised the need for the present systematic review, devised the rationale, defined the objectives, designed the protocol, conducted the literature search, screened eligible studies against the inclusion criteria, completed data extraction, undertook critical analysis and risk of bias assessments, undertook the synthesis and integration of results, designed the figures and tables, produced the main conclusions, wrote the report, and produced the reference list. SM, MAN and SR contributed to screening eligible studies against inclusion criteria if their relevance as per that criteria was ambiguous. SM contributed to study design and content, whilst providing feedback and suggestions for refining early drafts of the report. MAN and SR provided feedback and suggestions for refining the article's content and structure, whilst overseeing and supervising the project throughout.

Compliance with ethical standards

Conflict of interest The authors declare no competing interests.

References

1. Reznick RK, MacRae H. Teaching surgical skills-changes in the wind. N Engl J Med. 2006;355:2664–9.
2. Badakere A, Chhablani PP, Chandrasekharan A, Ali MH, Kekunnaya R. Comparison of pediatric cataract surgical techniques between pediatric ophthalmology consultants and fellows in training: a video-based analysis. J Pediatr Ophthalomol Strabismus. 2019;56:83–8.
3. Kim TS, O'Brien M, Zafar S, Hager GD, Sikder S, Vedula S. Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. Int J Comput Assist Radiol Surg. 2019;14:1097–105.
4. Turnbull AMJ, Lash SC. Confidence of ophthalmology specialist trainees in the management of posterior capsule rupture and vitreous loss. Eye. 2016;30:943–8.
5. Muttuvelu DV, Andersen CU. Cataract surgery education in member countries of the European Board of Ophthalmology. Can J Ophthalmol. 2016;51:207–11.
6. Mery CM, Greenberg JA, Patel A, Jaik NP. Teaching and assessing the ACGME competencies in surgical residency. Bull Am Coll Surg. 2008;93:39.
7. Yule S, Parker SH, Wilkinson J, McKinley A, MacDonald J, Neill A, et al. Coaching non-technical skills improves surgical residents' performance in a simulated operating room. J Surg Educ. 2015;72:1124–30.
8. Gawande AA, Zinner MJ, Studdert DM, Brennan TA. Analysis of errors reported by surgeons at three teaching hospitals. Surgery. 2003;133:614–21.
9. Yule S, Flin R, Paterson-Brown S, Maran N. Non-technical skills for surgeons in the operating room: a review of the literature. Surgery. 2006;139:140–9.
10. Steeples LR, Hingorani M, Flanagan D, Kelly SP. Wrong intraocular lens events—what lessons have we learned? A review of incidents reported to the National Reporting and Learning System: 2010–2014 versus 2003–2010. Eye. 2016;30:1049–55.
11. Simon JW, Ngo Y, Khan S, Strogatz D. Surgical confusions in ophthalmology. Arch Ophthalmol. 2007;125:1515–22.
12. Rogers GM, Oetting TA, Lee AG, Grignon C, Greenlee E, Johnson AT, et al. Impact of a structured surgical curriculum on

ophthalmic resident cataract surgery complication rates. J Cataract Refract Surg. 2009;35:1956–60.
13. Pena G, Altree M, Field J, Sainsbury D, Babidge W, Hewett P, et al. Nontechnical skills training for the operating room: A prospective study using simulation and didactic workshop. Surgery. 2015;158:300–9.
14. Frank JR, Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. Med Teach. 2007;29:642–7.
15. Scott DJ, Dunnington GL. The new ACS/APDS Skills Curriculum: moving the learning curve out of the operating room. J Gastrointest Surg. 2008;12:213–21.
16. Azuara-Blanco A, Reddy A, Wilkinson G, Flin R. Safe eye surgery: non-technical aspects. Eye. 2011;25:1109–11.
17. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Ioannidis JPA, Clarke M, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol. 2009;62:e1–34.
18. Oluwatayo JA. Validity and reliability issues in educational research. J Educ Soc Res. 2012;2:391–400.
19. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet. 2001;357:945–9.
20. Brunckhorst O, Aydin A, Abboudi H, Sahai A, Khan MS, Dasgupta P, et al. Simulation-based ureteroscopy training: a systematic review. J Surg Educ. 2015;72:135–43.
21. Van Der Vleuten CP. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ Theory Pract. 1996;1:41–67.
22. Carter FJ, Schijven MP, Aggarwal R, Grantcharov T, Francis NK, Hanna GB, et al. Consensus guidelines for validation of virtual reality surgical simulators. Surgical Endosc Other Interventional Tech. 2005;19:1523–32.
23. ICO. International Council of Ophthalmology Surgical Assessment Tool: ICO-OSCAR in English, Mandarin Chinese, Portuguese, Russian, Spanish, Thai, Vietnamese, and French. 2018. http://www.icoph.org/resources/230/Surgical-Assessment-Tool-ICO-OSCAR-in-English-and-Spanish.html.
24. Golnik KC, Beaver H, Gauba V, Lee AG, Mayorga E, Palis G, et al. Cataract surgical skill assessment. Ophthalmology. 2011;118:427.
25. Dean WH, Murray NL, Buchan JC, Golnik K, Kim MJ, Burton MJ. Ophthalmic simulated surgical competency assessment rubric for manual small-incision cataract surgery. J Cataract Refract Surg. 2019;45:1252–7.
26. Golnik C, Beaver H, Gauba V, Lee AG, Mayorga E, Palis G, et al. Development of a new valid, reliable, and internationally applicable assessment tool of residents' competence in ophthalmic surgery (an American Ophthalmological Society thesis). Trans Am Ophthalmol Soc. 2013;111:24–33.
27. Swaminathan M, Ramasubramanian S, Pilling R, Li J, Golnik K. ICO-OSCAR for pediatric cataract surgical skill assessment. J Am Assoc Pediatr Ophthalmol Strabismus. 2016;20:364–5.
28. Green CM, Salim S, Edward DP, Mudumbai RC, Golnik K. The ophthalmology surgical competency assessment rubric for trabeculectomy. J Glaucoma. 2017;26:805–9.
29. Golnik KC, Law JC, Ramasamy K, Mahmoud TH, Okonkwo ON, Singh J, et al. The ophthalmology surgical competency assessment rubric for vitrectomy. Retina. 2017;37:1797–804.
30. Golnik KC, Motley WW, Atilla H, Pilling R, Reddy A, Sharma P, et al. The ophthalmology surgical competency assessment rubric for strabismus surgery. J Am Assoc Pediatr Ophthalmol Strabismus. 2012;16:318–21.
31. Motley WW, Golnik KC, Anteby I, Atilla H, Gole GA, Murillo C, et al. Validity of ophthalmology surgical competency assessment rubric for strabismus surgery in resident training. J Am Assoc Pediatr Ophthalmol Strabismus. 2016;20:184–5.
32. Juniat V, Golnik KC, Bernardini FP, Cetinkaya A, Fay A, Mukherjee B, et al. The ophthalmology surgical competency assessment rubric (OSCAR) for anterior approach ptosis surgery. Orbit. 2018;37:401–4.
33. Golnik KC, Gauba V, Saleh GM, Collin R, Naik MN, Devoto M, et al. The ophthalmology surgical competency assessment rubric for lateral tarsal strip surgery. Ophthalmic Plast Reconstr Surg. 2012;28:350–4.
34. Grantcharov TP, Kristiansen VB, Bendix J, Bardram L, Rosenberg J, Funch-Jensen P. Randomized clinical trial of virtual reality simulation for laparoscopic skills training. Br J Surg. 2004;91:146–50.
35. Ezra DG, Aggarwal R, Michaelides M, Okhravi N, Verma S, Benjamin L, et al. Skills acquisition and assessment after a microsurgical skills course for ophthalmology residents. Ophthalmology. 2009;116:257–62.
36. Bann SD, Khan MS, Darzi AW. Measurement of surgical dexterity using motion analysis of simple bench tasks. World J Surg. 2003;27:390–4.
37. Saleh GM, Gauba V, Mitra A, Litwin AS, Chung AKK, Benjamin L. Objective structured assessment of cataract surgical skill. Arch Ophthalmol. 2007;125:363–6.
38. Thomsen ASS, Bach-Holm D, Kjaerbo H, Hojgaard-Olsen K, Subhi Y, Saleh GM, et al. Operating room performance improves after proficiency-based virtual reality cataract surgery training. Ophthalmology. 2017;124:524–31.
39. Selvander M, Asman P. Ready for OR or not? Human reader supplements Eyesi scoring in cataract surgical skills assessment. Clin Ophthalmol. 2013;7:1973–7.
40. Selvander M, Asman P. Cataract surgeons outperform medical students in Eyesi virtual reality cataract surgery: evidence for construct validity. Acta Ophthalmol. 2013;91:469–74.
41. Feldman BH, Geist CE. Assessing residents in phacoemulsification. Ophthalmology. 2007;114:1586.
42. Gauba V, Saleh GM, Goel S. Ophthalmic plastic surgical skills assessment tool. Ophthal Plast Reconstr Surg. 2008;24:43–6.
43. Pilling RF, Bradbury JA, Reddy AR. Strabismus surgical skills assessment tool: development of a surgical assessment tool for strabismus surgery training. Am J Ophthalmol. 2010;150:275–8.
44. Cremers SL, Lora AN, Ferrufino-Ponce ZK. Global rating assessment of skills in intraocular surgery (GRASIS). Ophthalmology. 2005;112:1655–60.
45. Lee AG, Greenlee E, Oetting TA, Beaver HA, Johnson AT, Boldt HC, et al. The Iowa ophthalmology wet laboratory curriculum for teaching and assessing cataract surgical competency. Ophthalmology. 2007;114:e21–6.
46. Gertsch KR, Kitzmann A, Larson SA, Olson RJ, Longmuir RA, Sacher BA, et al. Description and validation of a structured simulation curriculum for strabismus surgery. J Am Assoc Pediatr Ophthalmol Strabismus. 2015;19:3–5.
47. Zarei-Ghanavati M, Ghassemi H, Salabati M, Mahmoudzadeh R, Liu C, Daniell M, et al. A surgical skills assessment rubric for pterygium surgery. Ocul Surf. 2020;18:494–8.
48. Smith RJ, McCannel CA, Gordon LK, Hollander DA, Giaconi JA, Stelzner SK, et al. Evaluating teaching methods of cataract surgery: validation of an evaluation tool for assessing surgical technique of capsulorhexis. J Cataract Refract Surg. 2012;38:799–806.
49. Smith RJ, McCannel CA, Gordon LK, Hollander DA, Giaconi JA, Stelzner SK, et al. Evaluating teaching methods: validation of an evaluation tool for hydrodissection and phacoemulsification portions of cataract surgery. J Cataract Refract Surg. 2014;40:1506–13.
50. Saleh GM, Wawrzynski JR, Saha K, Smith P, Flanagan D, Hingorani M, et al. Feasibility of human factors immersive

simulation training in ophthalmology the london pilot. JAMA Ophthalmol. 2016;134:905–11.
51. Hull L, Arora S, Kassab E, Kneebone R, Sevdalis N. Observational teamwork assessment for surgery: Content validation and tool refinement. J Am Coll Surg. 2011;212:234–43.
52. Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D. Development of a rating system for surgeons' non-technical skills. Med Educ. 2006;40:1098–104.
53. Flin R, Patey R. Non-technical skills for anaesthetists: Developing and applying ANTS. Best Pract Res Clin Anaesthesiol. 2011;25:215–27.
54. Mishra A, Catchpole K, Mcculloch P. The Oxford NOTECHS system: reliability and validity of a tool for measuring teamwork behaviour In the operating theatre. BMJ Qual Saf. 2009;18:104–8.
55. Wood TC, Maqsood S, Zoutewelle S, Nanavaty MA, Rajak S. Development of the HUman Factors in intraoperative Ophthalmic Emergencies Scoring System (HUFOES) for non-technical skills in cataract surgery. Eye. 2020;5:1–9.
56. Tsagkataki M, Choudhary A. Mersey deanery ophthalmology trainees' views of the objective assessment of surgical and technical skills (OSATS) workplace-based assessment tool. Perspect Med Educ. 2013;2:21–7.
57. Spiteri A, Aggarwal R, Kersey T, Benjamin L, Darzi A, Bloom P. Phacoemulsification skills training and assessment. Br J Ophthalmol. 2010;94:536–41.
58. Jacobsen MF, Konge L, Bach-holm D, la Cour M, Holm L, Højgaard-Olsen K, et al. Correlation of virtual reality performance with real-life cataract surgery performance. J Cart Refract Surg. 2019;45:1246–51.
59. Benjamin L. Selection, teaching and training in ophthalmology. Clin Exp Ophthalmol. 2005;33:524–30.
60. Henderson BA, Ali R. Teaching and assessing competence in cataract surgery. Curr Opin Ophthalmol. 2007;18:27–31.
61. Wood TC, Raison N, Haldar S, Brunckhorst O, McIlhenny C, Dasgupta P, et al. Training tools for nontechnical skills for surgeons—a systematic review. J Surg Educ. 2017;74:548–78.