


# Validity Evidence for Procedure-specific Competence Assessment Tools in Orthopaedic Surgery: A Scoping Review

Yibo Li, MD 

Robert Chan, MD, MSc,  
FRCS 

Matthew R.G. Menon, MD,  
MHSc, FRCS

Joanna F. Ryan, MD

Brett Mador, MD, MHPE,  
FRCS

Sandra M. Campbell, MLS

Simon R. Turner, MD, MEd,  
FRCS

From the Department of Surgery, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Alberta, Canada (Dr. Li, Dr. Chan, Dr. Menon, Dr. Ryan, Dr. Mador, and Dr. Turner); the Western Upper Limb Facility, Sturgeon Community Hospital, St. Albert, Alberta, Canada (Dr. Chan); and the John W. Scott Health Sciences Library, University of Alberta, Edmonton, Canada (Ms. Campbell).

Correspondence to Dr. Li: [li4@ualberta.ca](mailto:li4@ualberta.ca)

None of the following authors or any immediate family member has received anything of value from or has stock or stock options held in a commercial company or institution related directly or indirectly to the subject of this article: Dr. Li, Dr. Chan, Dr. Menon, Dr. Ryan, Dr. Mador, Ms. Campbell, and Dr. Turner.

Disclosures: This work has previously been presented at the Canadian Orthopaedic Association Annual Meeting in 2022 (Quebec City, QC), at the Canadian Orthopaedic Residents' Association Annual Meeting in 2022 (Quebec City, QC), and at the University of Alberta Department of Surgery Tom Williams Research Day in 2022 (Edmonton, AB).

*JAAOS Glob Res Rev* 2024;8: e23.00065

DOI: 10.5435/JAOSGlobal-D-23-00065

Copyright © 2024 The Authors. Published by Wolters Kluwer Health, Inc. on behalf of the American Academy of Orthopaedic Surgeons. This is an open access article distributed under the Creative Commons Attribution-NoDerivatives License 4.0 (CC BY-ND) which allows for redistribution, commercial and non-commercial, as long as it is passed along unchanged and in whole, with credit to the author.

## ABSTRACT

**Introduction:** Competency-based training requires frequent assessment of residents' skills to determine clinical competence. This study reviews existing literature on procedure-specific competence assessment tools in orthopaedic surgery.

**Methods:** A systematic search of eight databases up to May 2023 was conducted. Two reviewers independently assessed validity evidence and educational utility of each assessment tool and evaluated studies' methodological quality.

**Results:** Database searching identified 2,556 unique studies for title and abstract screening. Full texts of 290 studies were reviewed; 17 studies met the inclusion criteria. Bibliography review identified another five studies, totaling 22 studies examining 24 assessment tools included in the analysis. These tools assessed various orthopaedic surgery procedures within trauma, sports medicine, spine, and upper extremity. Overall validity evidence was low across all studies, and was lowest for consequences and highest for content. Methodological quality of studies was moderate. Educational utility assessment was not explicitly done for most tools.

**Discussion:** The paucity of current procedure-specific assessment tools in orthopaedic surgery lacks the validity evidence required to be used reliably in high-stake summative assessments. Study strengths include robust methodology and use of an evidence-based validity evidence framework. Poor-quality existing evidence is a limitation and highlights the need for evidence-based tools across more subspecialties.

**C**hanges in orthopaedic surgery residency training brought on by work-hour restrictions and reduced surgical caseloads have resulted in programs incorporating new evaluation techniques of residents.<sup>1-3</sup> The assessment of technical skills in the age of competency-based medical

education relies on frequent evaluations by multiple observers over time and is turning from subjective toward objective assessments.<sup>1</sup>

Current objective assessment tools can be classified as global rating scales, procedure-specific tools, or hybrid scales.<sup>4</sup> Global rating scales are generic tools that can be used to assess performance for multiple different procedures, whereas procedure-specific tools can best address the specificity required for competency-based medical education and generate specific feedback for trainees.<sup>4</sup> Hybrid scales combine task-specific checklists with global rating scales and enjoy the benefits of both but as a result take longer to complete.<sup>4</sup>

Although numerous assessment tools in orthopaedic surgery have been developed,<sup>4</sup> the validity evidence supporting these tools is lacking.<sup>3,4</sup> Other surgical specialties including general surgery and cardiothoracic and vascular surgery have used a validity framework based on content, response process, internal structures, relation to other variables, and consequences to critically appraise assessment tools, with good interrater reliability.<sup>5-10</sup> Although other orthopaedic surgery assessment tools have been previously evaluated in the literature,<sup>3,4,11</sup> no review studies have specifically examined procedure-specific tools. The purpose of this study was to systematically review the literature on procedure-specific assessment tools in orthopaedic surgery and to assess validity evidence and educational utility for each tool, as well as to appraise the methodology of the identified studies. We hypothesize that there are few procedure-specific assessment tools supported by robust validity evidence.

## Methods

This study adhered to the Preferred Reporting Items for Systematic Review and Meta-analysis extension for Scoping Reviews.<sup>12</sup> The Preferred Reporting Items for Systematic Review and Meta-analysis extension for Scoping Reviews checklist is available in Supplemental Figure 1, <http://links.lww.com/JG9/A313>. A detailed description of the search methodology used has been reported elsewhere.<sup>9</sup>

### Search Strategy, Study Selection, and Data Extraction

A health sciences librarian conducted a systematic search in May 2023 on the following eight databases: OVID Medline, Ovid EMBASE, OVID PsycInfo, OVIDHealth and Psychosocial Instruments, SCOPUS, ProQuest

Dissertations and Theses Global, Cochrane Library, and PROSPERO. The concepts of ‘validation’ and ‘competence’ and ‘surgeons’ were used, and no limits were applied. Results were managed with the Covidence systematic review software. Reference lists of included studies were hand-searched for additional studies. At least two independent reviewers conducted initial title and abstract screening. Two reviewers (Y.L., R.C.) screened full-text articles. All conflicts were resolved by consensus decision. The inclusion criterion was assessment of validity evidence for procedure-specific orthopaedic surgery competency assessment instruments. Exclusion criteria were assessment of global rating scales (unless modified to be procedure-specific) or bedside procedures (eg, joint aspiration, closed reduction of a fracture, and physical examination), non-English studies, and conference abstracts and theses. Two reviewers (Y.L., R.C.) extracted information on each assessment tool using a Microsoft Excel (Microsoft Corp) template created by the authors at the beginning of the study (Appendix 1; <http://links.lww.com/JG9/A314>).

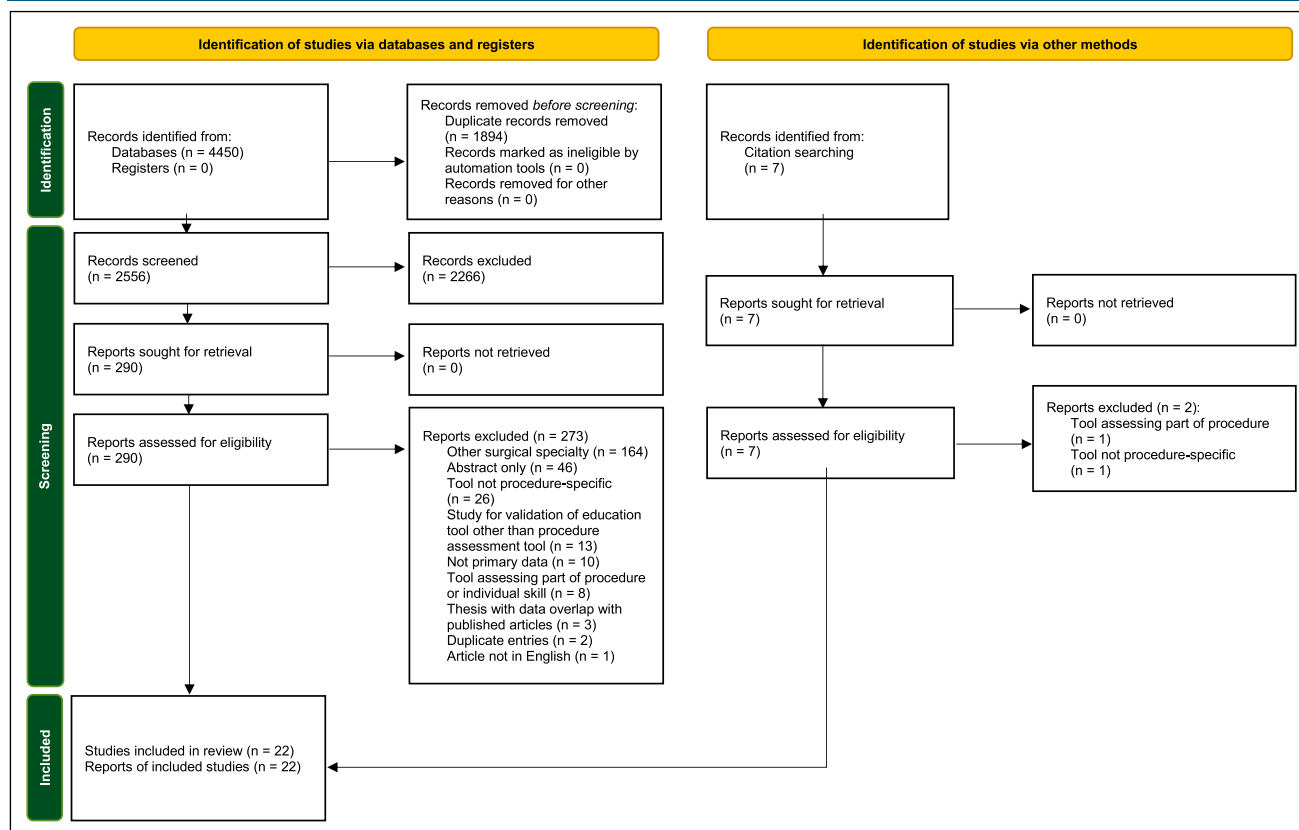
### Validity Evidence, Methodological Rigor, and Educational Utility Assessment

Two independent reviewers (Y.L., R.C.) assessed validity evidence, methodological rigor, and educational utility for each study. Disagreements were resolved by consensus decision. Validity evidence was scored using the five domains of the framework of Ghaderi et al<sup>8</sup> (content, response process, internal structure, relation to other variables, and consequences), with a maximum score of 15. Methodological rigor was assessed using the eight items of the Medical Education Research Study Quality Instrument framework, which assessed study design, sampling, type of data, data analysis, and outcome, with a maximum score of 18.<sup>13</sup> Educational utility was assessed using four domains of the Accreditation Council for Graduate Medical Education (ACGME) educational utility framework (ease of use, resources required, ease of interpretation, and educational impact).<sup>14</sup>

## Results

Database search identified 4,450 studies. After 1,894 duplicates were removed, 2,556 studies underwent title and abstract screening, excluding 2,266 studies. Full text of 290 studies were reviewed, and 17 studies met inclusion criteria (Figure 1). Additional review of reference

Figure 1



Preferred Reporting Items for Systematic Review and Meta-analysis flow diagram for study screening and inclusion. Reproduced with permission from Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71. doi: 10.1136/bmj.n71.

sections from these 17 studies identified another five studies meeting inclusion criteria, totaling 22 studies included in the analysis.

### Study and Assessment Tool Characteristics

We identified 22 studies using 24 procedure-specific surgical assessment tools (Table 1).<sup>15-36</sup> These tools assessed a variety of orthopaedic surgery procedures, including diagnostic knee arthroscopy and partial meniscectomy,<sup>21,31,36</sup> arthroscopic hip labral repair,<sup>26</sup> diagnostic shoulder arthroscopy,<sup>21,29</sup> arthroscopic rotator cuff repair and labral repair,<sup>15,16,19,22-25,27</sup> open surgical approaches to the shoulder (deltopectoral, lateral deltoid-splitting, and posterior),<sup>28</sup> shoulder arthroplasty,<sup>20</sup> arthroscopic hamstring anterior cruciate ligament reconstruction,<sup>30</sup> open carpal tunnel release,<sup>34,35</sup> trigger finger release,<sup>34</sup> percutaneous transforaminal endoscopic discectomy,<sup>17</sup> and fracture fixation.<sup>18,21,32-34</sup> All tools included a checklist of critical steps that were graded categorically. All tools except the Arthroscopic Bankart Metric and percutaneous

transforaminal endoscopic discectomy 10-step checklist were part of hybrid tools that also included a global rating scale.<sup>17,22-25</sup> All but five studies assessed tools in a simulation environment only; of the five, three assessed live operations<sup>21,29</sup> and two assessed arthroscopic recordings of operations.<sup>15,16</sup> Study participants included residents, fellows, and fellowship-trained attendings. Twenty-one tools were designed to evaluate resident performance, two tools were designed to distinguish between novice and experienced orthopaedic surgeons,<sup>15,16,22,25</sup> and one tool was designed to evaluate spine surgeons learning a new technique.<sup>17</sup> Only two of the studies studying four different tools specified that the tool was intended for formative assessment<sup>20,28</sup>; other studies did not distinguish whether the tool was meant for formative or summative assessment.

### Validity Evidence Assessment (Framework of Ghaderi et al)

Validity evidence was low across all studies, ranging from 1 to 9 of a maximum score of 15 (Table 2).

**Table 1. Studies Assessing Procedure-specific Surgical Assessment Tools in Orthopaedic Surgery**

Author	Year	Procedure	Setting	Number of Assessment Tools	Study Participants	Target Population	Formative/ Summative
Demirel	2022	Arthroscopic rotator cuff repair	Operating room (video recordings)	1	2 novice surgeons and 2 expert surgeons	Expert vs. novice surgeons	Not stated
Demirel	2017	Arthroscopic rotator cuff repair	Operating room (video recordings)	0 <sup>a</sup>	Expert surgeons (number not specified)	Surgeons	Not stated
Gadjradj	2022	Percutaneous transforaminal endoscopic discectomy	Operating room	1	Spine surgeons	Surgeons	Not stated
Hoyt	2022	Long bone open reduction and internal fixation	Simulation (animal model)	1	20 residents and attendings	Residents	Not stated
Hauschild	2021	Arthroscopic Bankart repair	Simulation (cadaver)	1	38 residents	Residents	Not stated
Lohre	2020	Reverse shoulder arthroplasty	Simulation (cadaver)	1	18 senior residents	Residents	Not stated
Wagner	2019	Shoulder arthroscopy, knee arthroscopy, ankle open reduction and internal fixation)	Operating room	3	8 residents in one study phase and 22 resident in subsequent study phase	Residents	Formative
Gallagher	2018	Arthroscopic Bankart repair	Simulation (video recordings of cadaver)	1	44 senior residents	Experienced vs. novice surgeons	Not stated
Angelo	2015 <sup>23</sup>	Arthroscopic Bankart repair	Simulation (video recordings of cadaver)	0 <sup>b</sup>	None	Experienced vs. novice surgeons	Not stated
Angelo	2015 <sup>24</sup>	Arthroscopic Bankart repair	Simulation (video recordings of cadaver)	0 <sup>b</sup>	12 senior residents and 10 shoulder surgeons	Experienced vs. novice surgeons	Not stated
Angelo	2015 <sup>25</sup>	Arthroscopic Bankart repair	Simulation (video recordings of dry model)	0 <sup>b</sup>	7 senior residents and 12 shoulder surgeons	Experienced vs. novice surgeons	Not stated

(continued)

**Table 1.** (continued)

Author	Year	Procedure	Setting	Number of Assessment Tools	Study Participants	Target Population	Formative/ Summative
Phillips	2017	Arthroscopic hip labral repair	Simulation (dry model)	1	37 residents, 5 sports medicine fellows, 5 attendings	Residents	Not stated
Dwyer	2017	Arthroscopic rotator cuff repair and labral repair	Simulation (dry model)	2	Rotator cuff repair: 39 residents, 7 sports medicine fellows, 5 sports medicine fellowship-trained attendings. Labral repair: 35 residents, 6 sports medicine fellows, 5 sports medicine fellowship-trained attendings. Labral repair: 35 residents, 6 sports medicine fellows, 5 sports medicine fellowship-trained attendings	Residents	Not Stated
Bernard	2016	3 open surgical approaches to shoulder (deltopectoral, lateral deltoid-splitting, posterior)	Simulation (cadaver)	3	23 residents	Residents	Not stated
Talbot	2015	Diagnostic shoulder arthroscopy	Operating room	1	6 residents	Residents	Formative
Dwyer	2015	Arthroscopic hamstring anterior cruciate ligament reconstruction	Simulation (dry model)	1	40 residents	Residents	Not stated
Cannon	2014	Diagnostic knee arthroscopy	Simulation (virtual simulator)	1	48 postgraduate year (PGY)-3 residents	Residents	Not stated
LeBlanc	2013	Ulnar fracture fixation	Simulation (virtual simulator and Sawbones)	1	22 residents	Residents	Not stated

(continued)

**Table 1.** (continued)

Author	Year	Procedure	Setting	Number of Assessment Tools	Study Participants	Target Population	Formative/ Summative
Yehyawi	2013	Complex tibial plafond articular fracture surgery	Simulation (dry model)	1	12 residents	Residents	Not stated
Van Heest	2012	Trigger finger release, open carpal tunnel release, and distal radius fracture fixation	Simulation (cadaver)	3	27 residents	Residents	Not stated
Van Heest	2009	Carpal tunnel release	Simulation (cadaver)	0 <sup>c</sup>	26 residents and 2 hand fellows	Residents	Not stated
Insel	2009	Diagnostic knee arthroscopy and partial meniscectomy	Simulation (cadaver)	1	59 residents, 3 sports medicine fellows, 6 sports medicine fellowship-trained attendings	Residents	Not stated

<sup>a</sup>This study evaluated the same tool as the other Demirel study.

<sup>b</sup>These studies all evaluated the same tool as the Gallagher study.

<sup>c</sup>This study evaluated one of the same tools as the other Van Heest study.

Overall, tools scored highest in the content validity domain. Three tools scored 3 (12.5%), 12 tools scored 2 (50.0%), and nine tools scored 1 (37.5%). A list of items was available for all but one tool.<sup>19</sup> All tools except five were developed by content experts (not specified in five tools).<sup>15-20</sup> Fourteen tools (58.3%) underwent the modified Delphi technique for revision.

Tools scored poorly in the response process domain. Four tools scored 2 (16.7%), eight tools scored 1 (33.3%), and 12 tools scored 0 (50.0%). Rater training (4/24, 16.7%), pilot testing (7/24, 29.2%), participant familiarity with the tool (3/24, 12.5%), and qualitative analysis of thought process (1/24, 4.2%) were sources of evidence in this category.

The internal structure domain scores were moderate, with nine tools scoring 2 (37.5%), 10 tools scoring 1 (41.7%), and five tools scoring 0 (20.8%). Most tools (19/24, 79.2%) were assessed by intertest reliability. Other forms of evidence presented included measures of interrater reliability (16/24, 66.7%), intrarater reliability (1/24, 4.2%), internal consistency (14/24, 58.3%), and item analysis (2/24, 8.3%).

Tools scored better in the relation to other variables domain, with five tools scoring the maximum of 3 (20.8%), seven tools scoring 2 (29.2%), nine tools scoring

1 (37.5%), and three tools scoring 0 (12.5%). Most tools were correlated with postgraduate level of training (18/24, 75.0%) and a global rating scale (12/24, 50.0%). Other variables correlated with the tools included pass/fail assessments (6/24, 25.0%), self-reported previous number of the assessed procedure performed (6/24, 25.0%), number of months spent in relevant subspecialty rotations (3/24, 12.5%), novice or expert status (1/24, 4.2%), knowledge test (1/24, 4.2%), and various other specialized tests (visualization scale, probing scale, and Precision Score, each 1/24, 4.2%).

Tools scored very poorly in the consequences domain, with three tools scoring 2 (12.5%), six tools scoring 1 (25.0%), and 15 tools scoring 0 (62.5%). Only one tool (4.2%) provided a cut score well supported by data, and only six tools (25.0%) demonstrated support from users for their educational utility and value as determined by postsurvey data.

### Methodological Quality (Medical Education Research Study Quality Instrument Framework)

Methodological quality of studies was moderate, with scores ranging from 5.5 to 16.5. Most studies scored 11.5 (6/22, 27.3%) or 12.5 of a maximum score of 18 (9/22,

**Table 2.** Detailed Validity Evidence for Procedure-specific Surgical Assessment Tools

Tool	Article(s)	Content (Max 3)	Response Process (Max 3)	Internal Structure (Max 3)	Relation to Other Variables (Max 3)	Consequences (Max 3)	Total Score (Max 15)
Arthroscopy rotator cuff repair metrics	Demirel 2017 and 2022	2	1	1	1	0	5
Percutaneous transforaminal endoscopic discectomy 10-step checklist	Gadraj 2022	1	0	0	0	0	1
OSATS checklist for long bone ORIF	Hoyt 2022	1	0	1	1	1	4
Procedure-specific checklist for arthroscopic Bankart repair	Hauschild 2021	1	0	0	0	0	1
OSATS checklist for reverse shoulder arthroplasty	Lohre 2020	1	1	0	0	0	2
Task-specific checklist for shoulder arthroscopy	Wagner 2019	2	2	1	1	2	8
Task-specific checklist for knee arthroscopy	Wagner 2019	2	2	1	1	2	8
Task-specific checklist for ankle ORIF	Wagner 2019	2	2	1	1	2	8
Arthroscopic Bankart Metric	Gallagher 2018, Angelo 2015 and 2015 and 2015	3	1	2	1	0	6
Task-specific checklist for arthroscopic hip labral repair	Phillips 2017	2	0	1	1	0	4
Task-specific checklist for arthroscopic rotator cuff repair	Dwyer 2017	2	1	2	3	0	8
Task-specific checklist for arthroscopic labral repair	Dwyer 2017	2	1	2	3	0	8

(continued)

**Table 2.** (continued)

<b>Tool</b>	<b>Article(s)</b>	<b>Content (Max 3)</b>	<b>Response Process (Max 3)</b>	<b>Internal Structure (Max 3)</b>	<b>Relation to Other Variables (Max 3)</b>	<b>Consequences (Max 3)</b>	<b>Total Score (Max 15)</b>
OSATS checklist for deltopectoral approach to shoulder	Bernard 2016	2	0	2	3	0	7
OSATS checklist for lateral deltoid-splitting approach to shoulder	Bernard 2016	2	0	2	3	0	7
OSATS checklist for posterior approach to shoulder	Bernard 2016	2	0	2	3	0	7
Shoulder Objective Practical Assessment Tool for diagnostic shoulder arthroscopy	Talbot 2015	3	1	2	2	1	9
Task-specific checklist for arthroscopic anterior cruciate ligament reconstruction	Dwyer 2015	2	1	2	2	0	7
Procedural checklist for diagnostic knee arthroscopy	Cannon 2014	3	2	1	2	0	8
OSATS checklist for ulnar fracture fixation	LeBlanc 2013	1	1	1	1	1	5
Procedure-specific checklist for complex tibial plafond articular fracture surgery	Yehyawwi 2013	1	0	0	1	0	2
OSATS checklist for carpal tunnel release	Van Heest 2012 and 2019	1	0	2	2	1	6
OSATS checklist for trigger finger release	Van Heest 2012	1	0	1	2	1	5
OSATS checklist for distal radius fixation	Van Heest 2012	1	0	1	2	1	5

(continued)



**Table 2.** (continued)

Tool	Article(s)	Content (Max 3)	Response Process (Max 3)	Internal Structure (Max 3)	Relation to Other Variables (Max 3)	Consequences (Max 3)	Total Score (Max 15)
Basic Arthroscopic Knee Skill Scoring System checklist for diagnostic knee arthroscopy and partial meniscectomy	Insel 2009	2	0	0	2	0	4

OSATS = Objective Structured Assessment of Technical Skills; ORIF = open reduction and internal fixation

40.9%) (Table 3). One study (4.5%) designed to assess face and content validity for the Arthroscopic Bankart Metric tool scored 5.5 because it did not assess implementation of the tool.<sup>14</sup> Most studies (20/22, 90.9%) lost points for study design because they were single-group cross-sectional studies, and all studies lost points for outcome because no studies assessed a change in physician behaviors or patient or healthcare outcomes after the use of the tool.

### Educational Utility (Accreditation Council for Graduate Medical Education Framework)

Nearly all studies qualified as easy to use in the course of daily clinical or teaching activity with minimal set-up required (Table 4). However, only two studies reported on the time required to complete the assessment: after the first 50 assessments, the Shoulder Objective Practical Assessment Tool averaged 2 minutes 27 seconds to complete (range, 1 minute 29 seconds to 3 minutes 13 seconds),<sup>29</sup> while in the qualitative interview examining the various task-specific checklists (including for shoulder arthroscopy, knee arthroscopy, and ankle open reduction and internal fixation) by Wagner et al.,<sup>21</sup> participants noted that tools took 5 to 15 minutes to complete. Only one tool required resources beyond the documentation tools,<sup>15,16</sup> whereas all tools were completed by an individual assessor. Only one study both reported and met criteria for training requirements for assessors not exceeding an hour: the Objective Structured Assessment of Technical Skills (OSATS) checklist for ulnar fracture fixation reported only 10 minutes of assessor training.<sup>32</sup> Three studies examining the same tool (Arthroscopic Bankart Metric) reported on resources required as an 8-hour in-person meeting for reviewer training, which did not fulfill ACGME standards for training requirements.<sup>23-25</sup> Three studies exam-

ining two tools also reported data on ease of interpretation by providing evidence-based cut scores: the Arthroscopic Bankart Metric and procedural checklist for diagnostic knee arthroscopy both provide individual interpretable scores,<sup>24,25,31</sup> fulfilling ACGME standards for interpretability of individual scores,<sup>14</sup> whereas no studies reported data on educational impact.

### Discussion

Competency-based medical education relies heavily on an elaborate and robust assessment system to evaluate resident performance and readiness for independent practice.<sup>37</sup> In the orthopaedic surgery literature, procedure-specific surgical assessment tools are one component of this assessment system which have not previously been specifically reviewed.

This study identified 22 studies using 24 procedure-specific surgical assessment tools. Thirteen tools evaluated arthroscopic procedures, and only five evaluated fracture osteosynthesis. Orthopaedic subspecialties represented include only trauma, sports medicine, spine, and upper extremity. Considering the breadth of orthopaedic procedures, trainees are required to perform satisfactorily, and there is a clear lack of procedure-specific assessment tools to assess resident performance.

In keeping with our hypothesis, our study identified extensive variability in the validity evidence supporting procedure-specific assessment tools in orthopaedic surgery, with no studies scoring highly. Methodological quality was moderate for almost all studies. Most tools had limited ease of interpretation, with only two tools supported by a validated cut-score.<sup>24,25,31</sup> The Arthroscopic Bankart Metric was evaluated in both a shoulder simulator setting and a cadaveric shoulder setting in

**Table 3. Medical Education Research Study Quality Instrument (MERSQI) Scores**

Study	Year	Study Design (Max 3)	Sampling Institutions (Max 1.5)	Sampling Response Rate (Max 1.5)	Types of Data (Max 3)	Validity Evidence (Max 3)	Data Sophistication (Max 2)	Data Analysis (Max 1)	Outcomes (Max 3)	Total Score (Max 16.5)
Demirel	2022	1	Not specified	n/a	3	1	2	1	1.5	9.5
Demirel	2017	1	Not specified	n/a	3	1	2	1	1.5	9.5
Gadjradj	2022	1	1.5	1.5	3	0	2	1	1.5	11.5
Hoyt	2022	1	1	0.5	3	0	2	1	1.5	10
Hauschild	2021	2	0.5	1.5	3	0	2	1	1.5	11.5
Lohre	2020	3	1.5	0.5	3	0	2	1	1.5	12.5
Wagner	2019	1	0.5	0.5	3	2	2	1	1.5	11.5
Gallagher	2018	1	1.5	0.5	3	1	2	1	1.5	11.5
Angelo	2015 <sup>23</sup>	1	1.5	n/a	1	1	0	0	1	5.5
Angelo	2015 <sup>24</sup>	1	1.5	0.5	3	2	2	1	1.5	12.5
Angelo	2015 <sup>25</sup>	1	1.5	0.5	3	2	2	1	1.5	12.5
Phillips	2017	1	0.5	0.5	3	2	2	1	1.5	11.5
Dwyer	2017	1	0.5	1.5	3	2	2	1	1.5	12.5
Bernard	2016	1	0.5	1.5	3	2	2	1	1.5	12.5
Talbot	2015	1	1.5	0.5	3	2	2	1	1.5	12.5
Dwyer	2015	1	0.5	0.5	3	2	2	1	1.5	11.5
Cannon	2014	3	1.5	1.5	3	3	2	1	1.5	16.5
LeBlanc	2013	2	0.5	1.5	3	2	2	1	1.5	13.5
Yehyaw	2013	1	0.5	0.5	3	1	2	1	1.5	10.5
Van Heest	2012	1	0.5	1.5	3	2	2	1	1.5	12.5
Van Heest	2009	1	0.5	1.5	3	2	2	1	1.5	12.5
Insel	2009	1	0.5	1.5	3	2	2	1	1.5	12.5

n/a = not applicable

**Table 4.** Accreditation Council for Graduate Medical Education Educational Utility Criteria

Tool	Article(s)	Ease of Use <sup>a</sup>	Resources Required <sup>b</sup>	Ease of Interpretation <sup>c</sup>	Educational Impact
Arthroscopy rotator cuff repair metrics	Demirel 2017 and 2022	N	N	N	N
Percutaneous transforaminal endoscopic discectomy 10-step checklist	Gadjradj 2022	1,2	1,3	1	N
OSATS checklist for long bone ORIF	Hoyt 2022	1,2	1,3	N	N
Procedure-specific checklist for arthroscopic Bankart repair	Hauschild 2021	1,2	1,3	N	N
OSATS checklist for reverse shoulder arthroplasty	Lohre 2020	1,2	1,3	N	N
Task-specific checklist for shoulder arthroscopy	Wagner 2019	1,2,3	1,3	N	N
Task-specific checklist for knee arthroscopy	Wagner 2019	1,2,3	1,3	N	N
Task-specific checklist for ankle ORIF	Wagner 2019	1,2,3	1,3	N	N
Arthroscopic Bankart Metric	Gallagher 2018, Angelo 2015 and 2015 and 2015	1,2	1,3	1	N
Task-specific checklist for arthroscopic hip labral repair	Phillips 2017	1,2	1,3	N	N
Task-specific checklist for arthroscopic rotator cuff repair	Dwyer 2017	1,2	1,3	N	N
Task-specific checklist for arthroscopic labral repair	Dwyer 2017	1,2	1,3	N	N
OSATS checklist for deltopectoral approach to shoulder	Bernard 2016	1,2	1,3	N	N
OSATS checklist for lateral deltoid-splitting approach to shoulder	Bernard 2016	1,2	1,3	N	N
OSATS checklist for posterior approach to shoulder	Bernard 2016	1,2	1,3	N	N
Shoulder Objective Practical Assessment Tool for diagnostic shoulder arthroscopy	Talbot 2015	1,2,3	1,3	N	N
Task-specific checklist for arthroscopic anterior cruciate ligament reconstruction	Dwyer 2015	1,2	1,3	N	N

(continued)

**Table 4.** (continued)

Tool	Article(s)	Ease of Use <sup>a</sup>	Resources Required <sup>b</sup>	Ease of Interpretation <sup>c</sup>	Educational Impact
Procedural checklist for diagnostic knee arthroscopy	Cannon 2014	1,2	1,3	N	N
OSATS checklist for ulnar fracture fixation	LeBlanc 2013	1,2	1,2,3	N	N
Procedure-specific checklist for complex tibial plafond articular fracture surgery	Yehyaw 2013	1,2	1,3	N	N
OSATS checklist for carpal tunnel release	Van Heest 2012 and 2019	1,2	1,3	N	N
OSATS checklist for trigger finger release	Van Heest 2012	1,2	1,3	N	N
OSATS checklist for distal radius fixation	Van Heest 2012	1,2	1,3	N	N
Basic Arthroscopic Knee Skill Scoring System checklist for diagnostic knee arthroscopy and partial meniscectomy	Insel 2009	1,2	1,3	N	N

<sup>a</sup>1 = The assessment tool is easily carried or accessed in the course of daily clinical or teaching activity, 2 = The tool requires little special set-up, 3 = The tool requires less than 20 minutes for the assessor to complete.

<sup>b</sup>1 = No additional resources are required beyond the documentation tools, 2 = Training requirements for assessors do not exceed an hour, 3 = No additional persons other than an individual assessor are required to complete the evaluation.

<sup>c</sup>1 = Individual scores are interpretable.

novice and experienced surgeons, and the mean performance of the experienced group was used to establish a benchmark for proficiency.<sup>24,25</sup> The procedural checklist for diagnostic knee arthroscopy by Cannon et al.<sup>31</sup> set a proficiency score of at least 83% based on the average proficiency score of five community-based orthopaedic surgeons. No studies commented on the educational impact of the assessment tools, suggesting the lack of trainee engagement in the development and evaluation of these tools. Furthermore, no studies evaluated the effect of use of tools on outcomes (including change in physician behaviors or patient or healthcare outcomes).

Our study is limited by the low-quality evidence identified and the heterogeneity of procedure-specific assessment tools identified, precluding any comparative analysis. Although previous studies have reviewed current assessment tools in orthopaedic surgery,<sup>3,4,11</sup> the strengths of our study lie in its robust methodology and its focus on procedure-specific assessment tools. The comprehensive search strategy used included eight databases to mitigate the risk of missing relevant pub-

lications. In addition, our study assessed validity evidence, methodological rigor, and educational utility using frameworks that have been previously shown to have good interrater reliability.<sup>8,38</sup>

A combination of assessment formats is required to provide a complete evaluation of trainee performance. The underlying differences in the structure and intent of each tool, however, do create challenges when comparing tools from different categories. This review is an attempt at comparing all described orthopaedic procedure-specific assessments.

Similar to the findings of this study, the literature of procedure-specific surgical assessment tools in general surgery and cardiovascular and thoracic (CVT) surgery shows insufficient representation of the breadth of procedures, with 23 general surgeries and eight CVT tools identified.<sup>9,10</sup> Analogous to our findings, most general surgery tools evaluated laparoscopic procedures.<sup>9</sup> Regarding validity evidence, general surgery and CVT tools had similarly stronger evidence of content validity and weaker evidence of response process and consequence, whereas orthopaedic surgery had stronger evidence in relation to other

variables.<sup>9,10</sup> Methodological rigor and educational utility were similar to the findings in this study.<sup>9,10</sup>

A clear future direction for research in this field is the development of procedure-specific assessment tools in orthopaedic subspecialties using the established validity guidelines and methodologies.

## Acknowledgments

The authors acknowledge Abigail White for her contributions to initial abstract and title screening of studies.

## References

- Nousiainen MT, McQueen SA, Hall J, et al.: Resident education in orthopaedic trauma: The future role of competency-based medical education. *The Bone Joint J* 2016;98-B:1320-1325.
- Moorthy K, Munz Y, Sarker S, Darzi A: Objective assessment of technical skills in surgery. *BMJ* 2003;327:1032-1037.
- Velazquez-Pimentel D, Stewart E, Trockels A, Achan P, Akhtar K, Vaghela KR: Global rating scales for the assessment of arthroscopic surgical skills: A systematic review. *Arthroscopy* 2020;36:1156-1173.
- James HK, Chapman AW, Pattison GTR, Fisher JD, Griffin DR: Analysis of tools used in assessing technical skills and operative competence in trauma and orthopaedic surgical training: A systematic review. *JBJs Rev* 2020;8:e1900167.
- Messick S: Foundations of validity: Meaning and consequences in psychological assessment. *ETS Res Rep Ser* 1993;1993:i-18.
- Messick S: Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol* 1995;50:741-749.
- Messick S: Consequences of test interpretation and sse: The fusion of validity and values in psychological assessment. *ETS Res Rep Ser* 1998;1998:i-32.
- Ghaderi I, Manji F, Park YS, et al.: Technical skills assessment toolbox: A review using the unitary framework of validity. *Ann Surg* 2015;261:251-262.
- Ryan JF, Mador B, Lai K, Campbell S, Hyakutake M, Turner SR: Validity evidence for procedure-specific competence assessment tools in general surgery: A scoping review. *Ann Surg* 2022;275:482-487.
- White A, Muller Moran HR, Ryan J, Mador B, Campbell S, Turner SR: Validity evidence for procedure-specific competency assessment tools in cardiovascular and thoracic surgery: A scoping review. *J Surg Educ* 2022; S1931720422000538.
- Middleton RM, Baldwin MJ, Akhtar K, Alvand A, Rees JL: Which global rating scale?: A comparison of the ASSET, BAKSSS, and IGARS for the assessment of simulated arthroscopic skills. *The J bone joint Surg Am volume* 2016;98:75-81.
- Page MJ, McKenzie JE, Bossuyt PM, et al.: The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.
- Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM: Association between funding and quality of published medical education research. *JAMA* 2007;298:1002-1009.
- Swing SR, Clyman SG, Holmboe ES, Williams RG: Advancing resident assessment in graduate medical education. *J Grad Med Educ* 2009;1: 278-286.
- Demirel D, Palmer B, Sundberg G, et al: Scoring metrics for assessing skills in arthroscopic rotator cuff repair: Performance comparison study of novice and expert surgeons. *Int J Comput Assist Radiol Surg* 2022;17:1823-1835.
- Demirel D, Yu A, Cooper-Baer S, et al: A hierarchical task analysis of shoulder arthroscopy for a virtual arthroscopic tear diagnosis and evaluation platform (VATDEP). *Int J Med Robotics Comput Assist Surg* 2017;13:e1799.
- Gadjradj PS, Schutte P, Vreeling A, Depauw P, Harhangi BS: Assessing the learning process of transforaminal endoscopic discectomy for sciatica. *Neurospine* 2022;19:563-570.
- Hoyt B, Clark D, Lundy A, Schroeder S, Wagner S, Langhammer C: Validation of a high-fidelity fracture fixation model for skill acquisition in orthopedic surgery residents. *J Surg Educ* 2022;79:1282-1294.
- Hauschild J, Rivera J, Johnson A, Burns T, Roach C: Shoulder arthroscopy simulator training improves surgical procedure performance: A controlled laboratory study. *Orthop J Sports Med* 2021;9: 23259671211003873.
- Lohre R, Bois AJ, Pollock JW, et al: Effectiveness of immersive virtual reality on orthopedic surgical skills and knowledge acquisition among senior surgical residents: A randomized clinical trial. *JAMA Netw Open* 2020;3:e2031217.
- Wagner N, Acai A, McQueen S, et al: Enhancing formative feedback in orthopaedic training: Development and implementation of a competency-based assessment framework. *J Surg Educ* 2019;76:1376-1401.
- Gallagher AG, Ryu RKN, Pedowitz RA, Henn P, Angelo RL: Inter-rater reliability for metrics scored in a binary fashion — performance assessment for an arthroscopic Bankart repair. *J arthroscopic Relat Surg* 2018;34: 2191-2198.
- Angelo RL, Ryu RKN, Pedowitz RA, Gallagher AG: Metric development for an arthroscopic Bankart procedure: Assessment of face and content validity. *J Arthroscopic Relat Surg* 2015;31:1430-1440.
- Angelo RL, Ryu RKN, Pedowitz RA, Gallagher AG: The Bankart performance metrics combined with a cadaveric shoulder create a precise and accurate assessment tool for measuring surgeon skill. *J Arthroscopic Relat Surg* 2015;31:1655-1670.
- Angelo RL, Pedowitz RA, Ryu RKN, Gallagher AG: The Bankart performance metrics combined with a shoulder model simulator create a precise and accurate training tool for measuring surgeon skill. *J Arthroscopic Relat Surg* 2015;31:1639-1654.
- Phillips L, Cheung J, Whelan D, et al: Validation of a dry model for assessing the performance of arthroscopic hip labral repair. *Am J Sports Med* 2017;45:2125-2130.
- Dwyer T, Schachar R, Leroux T, et al.: Performance assessment of arthroscopic rotator cuff repair and labral repair in a dry shoulder simulator. *J Arthroscopic Relat Surg* 2017;33:1310-1318.
- Bernard JA, Dattilo JR, Srikumaran U, Zikria BA, Jain A, LaPorte DM: Reliability and validity of 3 methods of assessing orthopedic resident skill in shoulder surgery. *J Surg Educ* 2016;73:1020-1025.
- Talbot CL, Holt EM, Gooding BWT, Tennent TD, Foden P: The shoulder Objective Practical Assessment Tool: Evaluation of a new tool assessing residents learning in diagnostic shoulder arthroscopy. *J Arthroscopic Relat Surg* 2015;31:1441-1449.
- Dwyer T, Slade Shantz J, Chahal J, et al.: Simulation of anterior cruciate ligament reconstruction in a dry model. *Am J Sports Med* 2015;43: 2997-3004.
- Cannon W, Garrett W, Hunter R, et al: Improving residency training in arthroscopic knee surgery with use of a virtual-reality simulator. A randomized blinded study. *J Bone Joint Surg Am* 2014;96:1798-1806.
- LeBlanc J, Hutchison C, Hu Y, Donnon T: A comparison of orthopaedic resident performance on surgical fixation of an ulnar fracture using virtual reality and synthetic models. *J Bone Joint Surg Am Vol* 2013;95:S1-S5.

33. Yehyawit T, Thomas T, Ohrt G, et al: A simulation trainer for complex articular fracture surgery. *J Bone Joint Surg Am* 2013;95:e92-e98.
34. VanHeest A, Kuzel B, Agel J, Putnam M, Kalliainen L, Fletcher J: Objective structured assessment of technical skill in upper extremity surgery. *J Hand Surg* 2012;37:332-337.
35. Van Heest A, Putnam M, Agel J, Shanedling J, McPherson S, Schmitz C: Assessment of technical skills of orthopaedic surgery residents performing open carpal tunnel release surgery. *J Bone Joint Surg Am Vol* 2009;91:2811-2817.
36. Insel A, Carofino B, Leger R, Arciero R, Mazzocca AD: The development of an objective model to assess arthroscopic performance. *J Bone Joint Surg Am Vol* 2009;91:2287-2295.
37. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR, for the International CBME Collaborators: The role of assessment in competency-based medical education. *Med Teach* 2010;32:676-682.
38. Cook DA, Reed DA: Appraising the quality of medical education research methods: The medical education research study quality Instrument and the Newcastle–Ottawa scale-education. *Acad Med* 2015; 90:1067-1076.
- Levels of evidence are described in the table of contents. In this article, references 3-4, 8, 11-12, 20, and 31-32 are level I studies. References 9-10, 13, 15, 17-19, 21-22, 24-25, 27-30, 34-36, and 38 are level II studies. References 16, 26, and 33 are level III studies. References 1, 14, 23, and 37 are level IV studies. References 2 and 5-7 is a level V report or expert opinion.