Advances in Health Sciences Education (2022) 27:659–689
https://doi.org/10.1007/s10459-022-10114-w

Examining the validity argument for the Ottawa Surgical Competency Operating Room Evaluation (OSCORE): a systematic review and narrative synthesis

Martha Spencer[1] · **Jonathan Sherbino**[1] · **Rose Hatala**[1]

Received: 15 March 2021 / Accepted: 2 April 2022 / Published online: 5 May 2022

Abstract
The Ottawa Surgical Competency Operating Room Evaluation (OSCORE) is an assessment tool that has gained prominence in postgraduate competency-based training programs. We undertook a systematic review and narrative synthesis to articulate the underlying validity argument in support of this tool. Although originally developed to assess readiness for independent performance of a procedure, contemporary implementation includes using the OSCORE for entrustment supervision decisions. We used systematic review methodology to search, identify, appraise and abstract relevant articles from 2005 to September 2020, across MEDLINE, EMBASE and Google Scholar databases. Nineteen original, English-language, quantitative or qualitative articles addressing the use of the OSCORE for health professionals' assessment were included. We organized and synthesized the validity evidence according to Kane's framework, articulating the validity argument and identifying evidence gaps. We demonstrate a reasonable validity argument for the OSCORE in surgical specialties, based on assessing surgical competence as readiness for independent performance for a given procedure, which relates to ad hoc, retrospective, entrustment supervision decisions. The scoring, generalization and extrapolation inferences are well-supported. However, there is a notable lack of implications evidence focused on the impact of the OSCORE on summative decision-making within surgical training programs. In non-surgical specialties, the interpretation/use argument for the OSCORE has not been clearly articulated. The OSCORE has been reduced to a single-item global rating scale, and there is limited validity evidence to support its use in workplace-based assessment. Widespread adoption of the OSCORE must be informed by concurrent data collection in more diverse settings and specialties.

Keywords Assessment · Medical education · Competency-based training · Entrustment rating scale · OSCORE

✉ Martha Spencer
mspencer@providencehealth.bc.ca

[1] The University of British Columbia, Vancouver, BC, Canada

Introduction

Competency-based medical education (CBME) is being adopted world-wide as a new approach to medical education, particularly in postgraduate training (Iobst et al., 2010). In many countries, including the Netherlands, USA and Canada, the shift to CBME has also come with the implementation of, and increased focus on, workplace-based assessment (WBA) as part of programmatic assessment. WBA uses low-stakes assessment tools, implemented in the authentic clinical environment, that are intended to encourage direct observation and feedback in an assessment for learning paradigm (Norcini et al., 2007).

A further innovation has been the introduction of WBAs that capture supervision judgements using entrustment supervision scales. Entrustment supervision scales are behaviourally-anchored rating scales that capture the level of supervision a learner requires to perform a clinical task as they progress towards unsupervised practice (Ten Cate, 2020). Entrustment supervision scales have been touted as having several benefits. Scales that anchor on the supervisor's perception of a trainee's progressive clinical ability should promote construct alignment between the rating and the priorities of the supervisor (Crossley et al., 2011). Entrustment anchors that closely align with the degree of supervision required during the clinical task may encourage supervisors to use the entire range of the scale when rating a performance. Supervisors who may have been reluctant to tell a resident they are "below average" using traditional rating scales may be more willing to record "I had to do it" if that accurately captures the supervision provided. Finally, entrustment supervision scales capitalize on the natural decision making of clinical supervisors, who decide daily whether learners can be allowed to undertake clinical tasks with or without supervision (Ten Cate, 2020).

While many different entrustment supervision scales have been developed, one prominent tool in use across North American residency training programs is the Ottawa Surgical Competency Operating Room Evaluation (OSCORE) (Gofton et al., 2012; Dudek et al., 2015; MacEwan et al., 2016; Ode et al., 2019; Thanawala et al., 2018; Saliken et al., 2019; Fitzpatrick et al., 2019; Cutrer et al., 2020; Dudek et al., 2019; Thanawala et al., 2019; Van Heest et al., 2019; Prudhomme et al., 2020; Gillis et al., 2020; Halman et al., 2020; Thoma et al., 2020; Meholick et al., 2020; RCPSC, 2021). In the Canadian postgraduate medical education (PGME) context, the OSCORE is promoted as a 'strongly recommended' WBA entrustment supervision tool. (RCPSC, 2021).

The OSCORE was originally developed as a tool that would allow surgical training programs to determine surgical residents' competence, defined as "readiness for independent performance of the particular procedure", in select procedures throughout the course of their training (Gofton et al., 2012, p. 1402). The OSCORE has 8 clinical items rated on a 5-point scale (1 = "I had to do it" to 5 = "I did not need to be there"), one yes/no question about ability to perform the procedure independently, and two open ended feedback questions (Gofton et al., 2012). The OSCORE is novel compared to other surgical evaluation tools; it assesses overall surgical competence instead of narrowly focusing on technical skill and it assesses a resident's ability to independently perform the procedure as opposed to comparing the resident with their peer group.

Although originally intended as an assessment of surgical procedure competence (Gofton et al., 2012), the OSCORE is currently utilized as an entrustment supervision scale as evidenced by its inclusion in a recent review on entrustment supervision scales (Ten Cate et al., 2020). While it makes conceptual sense that there would be a direct relationship between a supervisor's assessment of a resident's competence and the supervisor's level

of entrustment of the resident to perform the task independently, in reality entrustment is influenced by a host of factors and the relationship between competence, independence and entrustment is complex (Hauer et al., 2015; Gilchrist et al., 2021; Klasen and Lingard, 2021). Within surgical supervision, emerging evidence supports a relationship between a supervisor's assessment of competence and entrustment of an operative procedure (Ji et al., 2019). The confluence of a promotion of entrustment-based decisions within CBME with the language of the OSCORE anchors (Ten Cate et al., 2020) (e.g. "How much supervision did this trainee require to perform the procedure independently?") has influenced the OSCORE's evolution as an entrustment supervision scale.

While there have been multiple individual studies on the use of the OSCORE in medical education, no review has systematically examined them together to understand whether the OSCORE is measuring what it intends to measure, and its effect on learners and programs of assessment (i.e., the validity argument underlying the OSCORE). The frameworks for organizing validity arguments have evolved from the early categories of concept, criterion and construct validity to more unifying contemporary conceptualizations of validity in which all validity is construct validity, supported by different sources of evidence (Messick 1989). Kane's validity framework is one such contemporary validity framework which is highly versatile as it both highlights the sources of validity evidence and offers a framework for synthesizing that evidence into a validity argument (Kane, 2013). Kane's framework can be used for both quantitative and qualitative assessment tools, as well as quantitative and qualitative sources of validity evidence. Kane's framework has two major components, starting with the interpretation/use argument (IUA) for the assessment tool (i.e., explicitly articulating the decision being made about a learner). The IUA identifies the key assumptions and inferences associated with the assessment decision. Once the IUA has been articulated, the necessary and/or available evidence that tests the assumptions of the IUA is evaluated. Validity evidence is captured from multiple sources and categorized under one of four inferences: scoring (evidence that examines the translation of an observation to a score on the rating tool); generalization (evidence supporting the sampling and reliability of the measurement); extrapolation (what the score infers about real-world performance); and implications (the impact of the assessment on the learner, program and/or patient) (Kane, 2013).

In the current study, we address the gap in the literature between the individual studies and the overall validity argument for the OSCORE. We use systematic review methodology to gather validity evidence and Kane's framework to examine the validity argument, identifying strengths and weakness and potential areas for future research and development of the OSCORE. We address the question: What is the validity argument underlying the use of the OSCORE in assessing readiness for independent performance of a procedure by medical learners?

Methods

The methodology for this systematic review was based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocols (PRISMA-P) (Moher et al., 2015).

Search strategy

We searched MEDLINE, Embase and Google scholar from 2005 (the earliest papers on entrustment) to September 2020 with the assistance of a reference librarian. The initial search included terms related to assessment (competenc*, assess*, evaluat*, educational measure*), combined with "entrust*", and supplemented by searching 'OSCORE' as a text word in the databases. Additional studies were sought by hand-searching the reference lists from two published reviews of entrustment supervision scales (Rekman et al., 2016; Ten Cate et al., 2020).

Study inclusion and exclusion

We included original quantitative or qualitative full-text research studies published in English. Studies had to address the use of the OSCORE for assessment of health professionals. Modifications of the original tool were included, but new derivative tools (e.g. Ontario Bronchoscopy Assessment Tool, Ottawa Clinic Assessment Tool) were excluded. Health professionals included physicians, nurses, pharmacists, dentists, veterinarians, allied health professionals, medical lab technicians if they provided patient care, and clinical psychologists. Meeting abstracts were excluded.

All identified titles and abstracts, and subsequently full-text articles, were independently screened by two authors to identify those that met the inclusion criteria. Disagreements were resolved by consensus.

Data abstraction process

All articles included in the systematic review were reviewed for general study characteristics and sources of validity evidence as per Kane's validity framework (Kane, 2013). We followed a previously published guide to categorize the validity evidence under each of Kane's inferences (Cook et al., 2015). A data abstraction sheet was developed and used to record information relevant to assessment including: the clinical setting in which the OSCORE was used (health care profession, specialty, inpatient vs. outpatient, academic vs. community, geographical location, simulation vs. real life, medical vs. surgical specialty, procedural vs. clinical, academic vs. community), learner characteristics (level of training, number of learners, number of encounters/learner, voluntary vs mandatory participation, OSCORE learner training, incentives), assessor characteristics (title/rank, number of assessors, number of encounters/assessor, OSCORE training for rater) and study design (purpose of assessment, study duration, task evaluated, opportunities for feedback by participants).

The interpretation/use argument (IUA) was extracted if it was explicitly stated in the study. Sources of validity evidence were also extracted and organized using Kane's framework.

Study quality

Methodological quality of the included quantitative studies was appraised using the Medical Education Research Study Quality Instrument (MERSQI) (Reed et al., 2007).

Data synthesis

Two authors critically examined the extracted data and categorized the validity evidence, with discrepancies resolved by consensus. All authors contributed to data analysis and synthesis to articulate the overall validity argument for the OSCORE and identify evidence gaps.

Reflexivity

All of the authors either currently hold or have held educational leadership positions in postgraduate medical education related to assessment. Two of the authors (JS, RH) also have careers in education scholarship and RH has previously published using Kane's framework. While Kane's framework itself is not inherently associated with a specific philosophical position on assessment, it is helpful to articulate our philosophical positions as they will influence our examination of the validity evidence (Tavares et al., 2020). While we describe ourselves as holding predominantly post-positivist views on assessment of learning, we hold philosophical positions more closely aligned with constructivism for WBAs such as the OSCORE. Specifically, we view competence as demonstrated through authentic clinical tasks as interpersonal, co-constructed between learner, supervisor and patient, and socially situated with multiple dimensions.

Results

The initial search yielded 1491 articles that was narrowed, using the inclusion criteria, to 19 studies focused on the OSCORE (Fig. 1). Seventeen were quantitative studies and two were qualitative. Eighteen studies were in post-graduate medicine; one study was in undergraduate medicine. The majority of the post-graduate studies (13/18) were in surgical specialties, many of which included orthopedics residents (7) and general surgery residents (5). All of these surgical studies examined the original ($n=10$) or modified ($n=3$) multi-item OSCORE. The remaining post-graduate studies included emergency medicine (n = 2), internal medicine ($n=1$), critical care medicine ($n=1$), or multiple medical specialties ($n=1$). All of these non-surgical studies, and the undergraduate study, examined a single global rating scale (GRS) with the OSCORE entrustment anchors. Six studies used the OSCORE for assessment in a simulation setting.

The MERQSI scores for included quantitative studies ranged from 9 to 14 with a mean score of 11.7 out of a possible score of 18. We divided the MERQSI scores into terciles of methodological quality with 1–6.5 being low quality, 7–12.5 being moderate quality and 13–18 being high quality. The majority of quantitative studies (12/17) included in this review were of moderate methodological quality; the remaining studies (5/17) were of high methodological quality (Table 1). Table 1 summarizes each study, the MERSQI score, and the detailed validity evidence.

Below, we present a narrative summary of the validity evidence using Kane's framework. While not explicitly stated, the studies predominantly examined the OSCORE assessments through a post-positivist lens (e.g. describing minimizing rater bias or considering reliability as the gold-standard for generalizability). The results presented below are consistent with this post-positivist perspective.

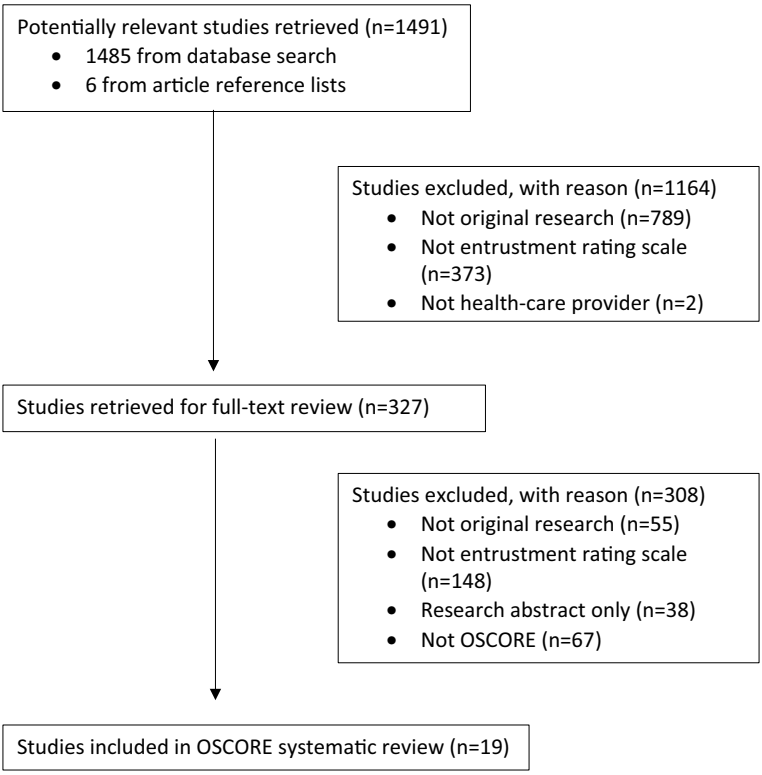

Fig. 1 PRISMA flow diagram

Interpretation/use argument (IUA)

The OSCORE was created as a "succinct surgical assessment tool that could be used to evaluate competence on *any* surgical procedure" (Gofton et al., 2012, p. 1402), where surgical competence was defined as "readiness for independent performance of a particular procedure" (Gofton et al., 2012, p. 1402). This IUA is consistent across the surgical postgraduate workplace-based studies included in our review. These surgical studies chose operative procedures across a range of different surgical specialties and assessed a resident's ability to perform a particular procedure independently using Gofton et al.'s five anchors. In the non-surgical postgraduate studies in which non-procedural skills were assessed, a clear IUA was not articulated although the studies imply an IUA of readiness for independent performance of a task. In the undergraduate study, assessors were asked to document the extent to which they had to intervene in the clinical task (Cutrer et al., 2020). By contrast, most studies in the simulation setting focused on assessing competence (Gerull et al., 2019; Halman et al., 2020; Prudhomme et al., 2020).

Validity argument: 1) Scoring

Evidence supporting the scoring inference describes how observation of performance is translated and captured as a numeric score or written comment (Cook et al., 2015). Only

Table 1 Description and detailed validity evidence for included studies

	Study description	Scoring	Generalization	Extrapolation	Implications	MERSQI score
Gofton et al., (2012)	Orthopedics + general surgery Operative cases Pilot with 20 residents assessed by 11 staff surgeons in orthopedics, then tool revised and studied with 37 residents (PGY1-5) assessed by 34 staff surgeons in orthopedics and general surgery	Development of OSCORE rating scale with descriptive anchors: original 14 items, piloted and revised to 9 items plus prompts for feedback Focus groups to examine response-process: colloquial anchors closely reflected real-world assessment Assessors found scale easy to use, as not norm-referenced but instead standard of independence Item-total correlation-0.57–0.82 for 8 items Rater training	Pre-specified 6 orthopedic and 5 general surgery procedures 163 procedures assessed G-coefficient 0.80 with 5 observations Largest proportion of variance due to trainees and observations (assessor nested in observation)	Expert-novice differences Correlation of 0.66 between total procedure OSCORE (average across 8-items) and yes/no independent item No difference based on procedural complexity	Residents found tool improved the amount and quality of feedback and helped to define the important aspects of the case Residents accepted low scores, as helped focus on which areas of surgical skill needed attention to become independent	13

Table 1 (continued)

	Study description	Scoring	Generalization	Extrapolation	Implications	MERSQI score
Dudek et al. (2015)	Orthopedics + general surgery + urology Operative cases Comparison of electronic version of the OSCORE [OSCORE (e)] vs. paper OSCORE 54 residents and 56 staff surgeons written assessment of acceptability of OSCORE(e)	OSCORE(e) beta-tested and piloted with local physicians prior to implementation Rater training	Pre-specified 6 orthopedic, 5 general surgery and 4 urology procedures 27 procedures assessed by OSCORE(e), historically 163 by OSCORE over similar timeframe		Feasibility: only 27/440 eligible OSCORE(e) cases were assessed Acceptability: Approx. 25% of respondents willing to do OSCORE but not OSCORE(e) OSCORE(e) barriers: technological factors, time, lost feedback opportunities as OSCORE(e) not completed in real-time with resident	9.5

Table 1 (continued)

	Study description	Scoring	Generalization	Extrapolation	Implications	MERSQI score
MacEwan et al. (2016)	Orthopedic surgery Benchtop simulation of orthopedic OSATS surgical case, videotaped and blinded + short-answer questions on treatment and diagnosis 21 participants (19 PGY-1 to PGY-5 residents and 2 staff surgeons) assessed by 2 different staff surgeons Same rater completed OSCORE and OSATS GRS and OSATS checklist	Rater training and assessment of training	Single simulated case, 21 assessments G-coefficient of 0.90 (8 items, 2 raters) Largest proportion of variance due to trainees and training level D-study, 1 rater, reliability would be 0.83 Inter-rater reliability of 0.89 for yes/no independent	Expert-novice differences Significant difference in total OSCORE for residents deemed competent to perform independently vs not OSCORE and OSATS GRS correlation of 0.96, no significant correlation with OSATS checklist		12

Table 1 (continued)

	Study description	Scoring	Generalization	Extrapolation	Implications	MERSQI score
Ode et al. (2019)	Orthopedic surgery Operative cases 19 residents (PGY2-5), 37 orthopedic surgeons electronic OSCORE		19 suggested orthopedic procedures, but any procedure accepted 326 assessments	Expert-novice differences for all levels except PGY3 vs. PGY4	Feasibility: residents requested mean 1.1 assessments/week (66.5% of eligible) 63.5% of requested assessments were completed 1/3 of residents found OSCORE form too long 77.8% of residents received immediate verbal feedback from staff 58% residents, 56% faculty found feedback generated by OSCORE helpful for resident training, encouraged real-time structured feedback	11.5

Table 1 (continued)

	Study description	Scoring	Generalization	Extrapolation	Implications	MERSQI score
Dudek et al. (2019)	Procedural and non-procedural specialties 12 residents (PGY1-6) and 10 faculty at single institution Semi-structured interviews exploring OSCORE entrustment anchors vs. traditional work-based assessment anchors	Entrustment anchors felt to be concrete, transparent and align with training outcomes Residents perceived OSCORE lacked information on performance compared to peers and expected rate of progress, compared to traditional anchors Scale useful in procedural and outpatient contexts, with both junior and senior learners			Residents more accepting of low scores compared to traditional anchors OSCORE didn't change amount of direct observation by faculty	N/A as qualitative study

Table 1 (continued)

	Study description	Scoring	Generalization	Extrapolation	Implications	MERSQI score
Fitzpatrick et al. (2019)	Urology Operative cases 17 residents (PGY1-5) and 12 staff surgeons completed survey about use of OSCORE Electronic OSCORE coupled with surgical log	50% residents and 42% faculty found written comments most valuable aspect of OSCORE 0% residents and 17% faculty found 8 individual items most valuable 19% of residents and 17% of faculty found the 'yes/no' question most valuable		69% residents and 75% faculty thought OSCORE was a true indicator of their surgical skill	63% residents logged > 80% of eligible cases 75% of faculty where either willing or very willing to complete O-score evaluations when requested 75% residents thought verbal feedback was unchanged with OSCORE compared to previous, 25% residents and 50% faculty thought it increased 81% residents and 75% faculty felt OSCORE + case log had positive effect on training	9

Table 1 (continued)

	Study description	Scoring	Generalization	Extrapolation	Implications	MERSQI score
Saliken et al. (2019)	Orthopedic surgery Benchtop simulation of orthopedic OSATS surgical case, videotaped and blinded + addition of short-answer questions on treatment and diagnosis Compared single assessor rating single-item yes/no independent question to 9-item paper-based OSCORE rating 6–8 weeks later 21 residents (PGY1-5) assessed by 1 staff surgeon and 2 fellows	Potential leniency bias towards using top-end of scale Rater training for 2/3 raters	Single simulated case, 21 assessments Generalizability study: Inter-rater reliability 0.83 Internal consistency 0.89	Expert-novice differences Correlation of 0.72–0.87 between single-item and total procedure OSCORE, measured at two separate time-points		12.5
Thanawala et al. (2019)	General surgery Operative cases 67 residents (PGY1-5), 33 staff surgeons, 4 residency programs Compared OSCORE to other performance measures including OPRS by same rater, using electronic platform		1230 operative procedures assessed, not pre-specified	Expert-novice differences Correlation 0.84 between total procedure OSCORE and OPRS	Feasibility: OSCORE took < 2 min to complete, with comments	12

Table 1 (continued)

	Study description	Scoring	Generalization	Extrapolation	Implications	MERSQI score
Van Heest et al. (2019)	Orthopedic surgery Operative cases 294 residents (PGY1-5), 370 orthopedic surgeons, 16 residency programs Compared P-score (single item, global rating scale, summative assessment of competence at end of procedure) to OSCORE using cross-over design, both electronic	46% residents preferred P-score over OSCORE and vice-versa for faculty Rater training	25 pre-specified milestone operative procedures 1150 OSCORE assessments	Expert-novice differences for all levels except PGY-3 vs. PGY-4	Feasibility: residents requested OSCORE on 88% of eligible cases Faculty completed > 90% of resident requests > 70% of faculty thought OSCORE discriminates between residents faculty and residents felt OSCORE defined relevant skills to safely perform operation	14
Gillis et al. (2020)	Orthopedic surgery Simulation (cadavers) 34 residents (PGY3-5) assessed by orthopedic staff (number not specified) Same rater completed OSCORE and OSCE score, as part of 4 station OSCE	No difference in community-based vs university-based faculty ratings Rater training	11 simulated, pre-specified orthopedic procedures 96 assessments over 3 years	Expert-novice differences Correlation between OSCORE and OSCE score of 0.89		10.5

Table 1 (continued)

	Study description	Scoring	Generalization	Extrapolation	Implications	MERSQI score
Modified OSCORE						
Thanawala et al. (2018)	Surgery (unspecified) Operative cases 43 residents, 23 surgeons logging cases and assessments on a new electronic platform Compared timeliness of staff completing modified OSCORE (12 items not specified) to timeliness of completing end-of-rotation evaluations		358 operative cases assessed, not pre-specified, for OSCOREs (610 end-of-rotation evaluations)		Feasibility: median lag of 1 day for completing OSCOREs compared to 35 days for completing end-of-rotation 75% faculty complete OSCORE within 2 min	9

Table 1 (continued)

	Study description	Scoring	Generalization	Extrapolation	Implications	MERSQI score
Gerull et al. (2019)	Multiple surgical specialties (general surgery, urology, obstetrics/gynecology) Robot-assisted laparoscopic surgical case 31 residents assessed by staff surgeons (number not specified) ROSCORE (robotics adapted OSCORE, 8 items) assessed pre- and post- robotics simulator training intervention by same assessor and self-assessment of NTLX (mental workload)	Potential rater bias as surgeon unblinded as to whether resident was pre- or post-robotics simulator training	62 operative cases	ROSCOREs improved post-training while NTLX mental workload score went down		9.5

Table 1 (continued)

	Study description	Scoring	Generalization	Extrapolation	Implications	MERSQI score
Meholick et al. (2020)	General surgery Laparoscopic operative cases 36 residents (PGY2-4) over 3 years Focus on 4/36 lower-performing residents requiring supplemental simulation training Modified, electronic OSCORE (12 items, rating anchors shortened) completed by faculty (number not specified)	Rater training	369 OSCORE assessments, 54 for the 4 lower-performing residents	Expert-novice differences; learning curves demonstrate increasing OSCORE with increasing operative exposure OSCORE learning curve performance lag for lower-performing residents OSCORE increases after simulation training for lower-performing residents	Feasibility: only 11% of eligible cases had an OSCORE assessment OSCORE was one of multiple assessments used to identify lower-performing residents	12.5

Table 1 (continued)

	Study description	Scoring	Generalization	Extrapolation	Implications	MERSQI score
Single GRS OSCORE						
Cutrer et al. (2020)	Undergraduate medicine 35 medical students, 94 faculty and residents as raters Ratings across 4 EPAs during post-clerkship acting internship Same rater completed Ottawa Co-activity scale (modified OSCORE with single item 4-anchor global rating scale) and Chen Supervisory scale Interviews with 3 faculty and 5 residentassessors	OSCORE modified by an internal assessment committee offaculty experts, edited anchors for clarity and brevity Intentional lack of rater training Raters felt Ottawa Co-activity scale retrospectively assessed amount of autonomy the student had been allowed Ottawa Co-activity scale may underestimate student ability, as organizational or supervisor factors may influence ratings Scale may be best suited for contexts where independence would not be allowed and when supervisor and student work closely together	4 pre-specified EPAs, 212 assessments; mean of 6/student	Strongest relationship between highest rating for both scales McNemar test indicates two scales are not the same	Scale anchors provided some guidance to faculty and students regarding expectations Assessors concerned about impact of ratings and comments on students' progress and future opportunities Comments felt to be critical to the development of the learner	13.5

Table 1 (continued)

	Study description	Scoring	Generalization	Extrapolation	Implications	MERSQI score
Lord et al. (2019)	Central venous catheter insertion (CVC) Medical and surgical specialties on ICU rotation (majority Family Practice) Simulation, videotaped 55 residents (PGY1-3) rated by 3 intensivists for each of two pre-specified CVC insertions Same rater completed single-item OSCORE (GRS) + itemized checklist + critical error tool + OSATS	Rater training and assessment of training (small error variance due to raters)	53 OSCORE assessments for each CVC insertion Dependability coefficients for OSCORE 0.78 (subclavian), 0.85 (internal jugular) G-study, major source of variance is resident D-study, reducing to 2 raters maintains dependability coefficient above 0.7			12
Martin et al. (2020)	GIM, anaesthesia, nephrology, EM 17 residents (PGY1-5) Focus group interviews examining influence of OSCORE entrustment ratings on feedback and assessment Single item OSCORE	Residents didn't perceive 5-point scale as offering different information than previous competency scales Residents perceived lack of information on expected rate of progress			Use of OSCORE impacted sense of self-efficacy Entrustment language, focused on autonomy, may reinforce 'performance culture' over growth mindset seeking observation and feedback	N/A as qualitative study

Table 1 (continued)

	Study description	Scoring	Generalization	Extrapolation	Implications	MERSQI score
Halman et al. (2020)	Internal medicine 91 residents (PGY 1–4), 42 faculty from 15 divisions of medicine at one institution 9-station OSCE progress test Single-item OSCORE (GRS) added to global rating scale (GRS) and training level rating scale (TLRS) and case-specific checklist (CL), completed by same rater	Item-total correlations (compared to overall exam score) of 0.30–0.79, depending on station	9 station OSCE OSCORE inter-station reliability of 0.79	Expert-novice differences Mean OSCORE correlation with PGY level of 0.65 Mean OSCORE correlation with pass/fail status of 0.67 all rating instruments highly correlated (0.85–0.93)	Using OSCORE for standard setting would have failed 38 residents (compared to 20 residents who actually failed exam) Acceptability- most examiners strongly agreed that the OSCORE and GRS best reflected trainee ability, and preferred them to the other ratings 62% examiners either comfortable or very comfortable with making entrustment decisions during an OSCE, 11% uncomfortable	12.5

Table 1 (continued)

	Study description	Scoring	Generalization	Extrapolation	Implications	MERSQI score
Prudhomme et al. (2020)	Emergency medicine Workplace vs. simulation-based assessment 9 PGY-1 residents, 10 faculty (single faculty assessor for workplace, two independent assessors for simulation) Single-item OSCORE (GRS) for one specific EPA		104 workplace assessments, 36 simulation assessments (mean 12 clinical and 4 simulation/resident) 1 pre-specified EPA Interrater reliability of simulation assessments of 0.86 G-coefficient of 0.35 for workplace, 0.75 for simulation D-study: 33 workplace assessments and 3 simulations to achieve reliability of 0.6	No significant correlation between workplace and simulation assessment (correlation coefficient 0.01) OSCOREs improve over time Mean workplace OSCOREs higher than simulation OSCOREs		14

Table 1 (continued)

	Study description	Scoring	Generalization	Extrapolation	Implications	MERSQI score
Thoma et al. (2020)	Emergency medicine 68 residents (PGY1), from 15 residency programs Single-item OSCORE (GRS)	< 10% EPAs rated at lower end of rating scale	28 pre-specified EPAs, organized across 4 stages of training (but 4th stage excluded as minimal data) 9842 assessments		All residents promoted to second stage of training, but 1/3 took longer than expected 3 months; over 60% of residents not promoted to 3rd stage at expected end of first year Substantial variability in number of assessments informing promotions decisions across programs, suggesting variability in competence committee standards	13

the original high methodological quality OSCORE study describes how the scale was developed (Gofton et al., 2012). An expert group referenced previously validated surgical assessment tools and created the unique entrustment supervision scale anchored with colloquial language that surgeons used to describe a resident's participation in a given procedure. Local surgeons reviewed the wording for relevance. The tool was piloted with orthopedic and general surgery residents and subsequently revised to its final form (Gofton et al., 2012).

None of the moderate methodological quality studies (MERSQI scores 9–12.5) that modified the OSCORE items (Gerull et al., 2019; Meholick et al., 2020; Thanawala et al., 2018) or reduced the OSCORE to a single GRS (Cutrer et al., 2020; Halman et al., 2020; Lord et al., 2019; MacEwan et al., 2016; Prudhomme et al., 2020; Thoma et al., 2020) described the development process for their modified scales.

Regarding response process, focus group participants in the original study (Gofton et al., 2012) felt the language of the anchors closely reflected real-world assessment. Another study reported that residents and faculty found the OSCORE anchors useful in procedural and non-procedural contexts for both junior and senior learners (Dudek et al., 2019). However, residents in a qualitative study did not perceive the single GRS OSCORE anchors as being different from traditional scales (Martin et al., 2020). Some residents preferred traditional anchors, which allowed comparison with their peers and gave them information on their expected rate of progress (Martin et al., 2020).

While rater training was undertaken in eight studies (Dudek et al., 2015; Gillis et al., 2020; Gofton et al., 2012; Halman et al., 2020; Lord et al., 2019; MacEwan et al., 2016; Meholick et al., 2020; Van Heest et al., 2019), none provided a detailed description and only one study was of high methodological quality (Gofton et al., 2012).

The influence of raters on scoring remains uncertain and was only assessed in three simulation studies of moderate methodological quality. In one unblinded simulation study that included rater training (Gillis et al., 2020), there were no differences between community faculty (less familiar with residents) versus academic faculty ratings, which the authors' suggest indicates minimal rater bias. Two of the simulation studies blinded the rater by using video-taped surgical procedures focusing only on the resident's gloved hands (MacEwan et al., 2016; Saliken et al., 2019). For studies in the clinical environment, raters were familiar with their residents.

In terms of entrustment scores, there is a tendency towards range restriction favouring the high end of the scale (Gofton et al., 2012; Saliken et al., 2019). The low end of the scale is infrequently used, even for very junior residents (Gofton et al., 2012). Less than 10% of first year emergency medicine residents scored 1 or 2 on the OSCORE, whereas greater than 60% rated 4 or 5 in a high methodological quality study (Thoma et al., 2020).

For psychometrics related to the scoring inference, the original OSCORE study found item-total correlations of 0.57–0.82 (Gofton et al., 2012) for 8 items. One moderate methodological quality OSCE-based study (Halman et al., 2020) found single OSCORE GRS item-total correlations of 0.30–0.79 by station compared to overall exam score.

Validity argument: 2) Generalization

The two major sources of evidence supporting the generalization inference are sampling and reliability (Cook et al., 2015). As outlined in Table 1, most studies took place at a single academic institution. A wide range of surgical procedures were assessed across moderate and high methodological quality studies. Seven studies included orthopedic procedures

(Gofton et al., 2012; Dudek et al., 2015; MacEwan et al., 2016; Ode et al., 2019; Saliken et al., 2019; Van Heest et al., 2019; Gillis et al., 2020), five studies included general surgery procedures (Gofton et al., 2012; Dudek et al., 2015; Thanawala et al., 2019; Gerull et al., 2019; Meholick et al., 2020), three studies included urological procedures (Dudek et al., 2015; Fitzpatrick et al., 2019; Gerull et al., 2019) and one included gynecological procedures (Gerull et al., 2019). The moderate methodological quality simulation studies examined only one type of procedure (Lord et al., 2019; MacEwan et al., 2016; Saliken et al., 2019). One study including surgical residents did not specify the types of procedures included (Thanawala et al., 2018) while one qualitative study included procedural and non-procedural specialties but did not specify which procedural specialties were included (Dudek et al., 2019). Among the non-surgical studies, two high methodological quality studies were in emergency medicine (Prudhomme et al., 2020; Thoma et al., 2020), one in internal medicine (Halman et al., 2020) and one in undergraduate medicine (Cutrer et al., 2020). In Gofton et al.'s original study, which included both orthopedic and general surgical residents and faculty, specialty accounted for little variability in item ratings, but this was not re-examined in the other studies (Gofton et al., 2012). All of these sampling issues limit generalizability to broader contexts, particularly non-surgical settings.

Examining the psychometric data, five studies employed generalizability theory to examine different sources of measurement error; two were of high methodological quality and three were of moderate quality (Gofton et al., 2012; MacEwan et al., 2016; Lord et al., 2018; Saliken et al., 2019; Prudhomme et al., 2020). Consistent across studies, variance attributed to raters was relatively small compared to variance attributed to residents (Gofton et al., 2012; Lord et al., 2019; MacEwan et al., 2016). High reliability was achievable with multiple assessments, ranging from a g-coefficient of 0.80 for five assessments by one rater in a clinical setting (Gofton et al., 2012) to 0.90 for eight assessments by two raters in a simulated context (MacEwan et al., 2016). In a high methodological quality study comparing workplace-based to simulation-based single GRS OSCORE assessments of first year residents' emergency resuscitation, the g-coefficient was markedly lower for the clinical setting (0.35 across twelve assessments with a single rater in the clinical setting compared to 0.75 across four cases and two raters in a simulated environment) (Prudhomme et al., 2020). This study suggests 33 work-based assessments would be required to achieve a reliability of 0.6. D-studies confirm that the number of raters can be reduced to one to two in the simulation setting without significantly impacting reliability (Lord et al., 2019; MacEwan et al., 2016).

Validity argument: 3) Extrapolation

Evidence supporting the extrapolation inference examines how performance on the OSCORE is related to real-world performance (Cook et al., 2015). As is evident in Table 1, the dominant source of evidence collected under the extrapolation inference is novice-expert differences. All ten of the moderate to high methodological quality studies (Gofton et al., 2012; MacEwan et al., 2016; Saliken et al., 2019; Van Heest et al., 2019; Thanawala et al., 2019; Ode et al., 2019; Prudhomme et al., 2020; Halman et al., 2020; Meholick et al., 2020; Gillis et al., 2020) that examined this relationship found that the OSCORE increased with level of training, even across months of training for first year emergency medicine residents (Prudhomme et al., 2020). Most of these studies are confounded by raters being unblinded to the resident level of training.

Additional extrapolation evidence, from both moderate and high methodological quality studies, examined the relationship of the OSCORE to other assessment tools. There are high correlations between the OSCORE and other surgical technical assessments completed at the same time (Gillis et al., 2020; MacEwan et al., 2016; Thanawala et al., 2019). There is moderate correlation between the single statement "resident competent to independently complete the procedure?" (Gofton et al., 2012 p. 1407) and the mean OSCORE rating (Gofton et al., 2012; Saliken et al., 2019) raising the possibility that the single-item score and the multi-component OSCORE function similarly. The OSCORE performed equivalently to the P-score, a single question summative assessment, in discriminating between levels of training (Van Heest et al., 2019).

Examining non-surgical assessments, Halman et al. found a high correlation between the OSCORE entrustment anchors and multiple other performance measures including a case-specific checklist, a GRS and a training level rating scale during an internal medicine OSCE in a moderate methodological quality study (Halman et al., 2020). In an undergraduate medicine, high methodological quality study, the OSCORE entrustment ratings were concordant with the higher ratings from another entrustment supervision scale (the Chen Supervisory scale), but mismatches were found for mid-range scores (Cutrer et al., 2020).

Only one high methodological quality study examined the relationship between simulation-based performance and clinical performance. There was no correlation (concordance correlation coefficient = -0.01, 95% CI -0.31–0.29, $p = 0.93$) between simulation and workplace-based single GRS OSCOREs for emergency medicine residents' emergency resuscitation with significantly higher scores in the workplace setting (Prudhomme et al., 2020).

Validity argument: 4) Implications

The implications inference addresses how available evidence impacts the learner, the faculty, the training program and/or patients and society (Cook et al., 2015). As demonstrated in Table 1, the vast majority of implications evidence focused on perceived feasibility of the OSCORE in practice and acceptability amongst staff and residents, across predominantly moderate methodological quality studies. Within the surgical studies, the OSCORE was generally found to be feasible for workplace-based assessment, using a feasibility standard of completing more than 50% of eligible assessments (Fitzpatrick et al., 2019; Ode et al., 2019; Van Heest et al., 2019). However, two studies found contrasting results with the studies by Meholick et al. and Dudek et al. reporting 11% and 6% OSCORE completion rates, respectively (Dudek et al., 2015, Meholick et al., 2020). Facilitators of high completion rates included email reminders, setting completion rate targets, and providing residents with immediate access to the OSCORE (Thanawala et al., 2018; Van Heest et al., 2019). The identified barriers included accessing electronic platforms, residents perceiving they were intruding on faculty, residents selectively choosing cases for assessment, and lack of time (Dudek et al., 2015; Ode et al., 2019). Two studies reported it took less than two minutes to complete the OSCORE after a surgical procedure (Thanawala et al., 2018, 2019), but in a third study, surgical residents felt it took too long to complete (Ode et al., 2019).

There was mixed evidence regarding acceptability of the OSCORE to residents and faculty in surgical specialties (Fitzpatrick et al., 2019; Ode et al., 2019; Van Heest et al., 2019). Across high methodological quality studies, use of the OSCORE helped to define important aspects of a surgical case for residents (Gofton et al., 2012; Van Heest et al., 2019) or clarified performance expectations (Cutrer et al., 2020). In two studies, residents

reported that they were more accepting of lower entrustment scores compared to traditional anchors (Dudek et al., 2019; Gofton et al., 2012) as the entrustment anchors highlighted performance deficits. In an internal medicine OSCE, faculty perceived the single GRS OSCORE entrustment scale to be a better measure of a resident's abilities than a global rating scale, training level rating scale or case-specific checklist (Halman et al., 2020).

A number of moderate and high methodological quality studies examined the impact of the OSCORE on feedback with surgical studies reporting that the amount and quality of verbal feedback improved (Fitzpatrick et al., 2019; Gofton et al., 2012; Ode et al., 2019; Van Heest et al., 2019). By contrast, in one study of residents across mixed specialties, residents did not find that direct observation increased with implementation of OSCORE assessments (Dudek et al., 2019). In a qualitative study with medicine subspecialty and emergency medicine residents, the OSCORE was felt to negatively impact residents' sense of self-efficacy and to potentially reinforce a performance mindset (seeking only positive assessments) over a growth mindset (seeking feedback to improve performance) (Martin et al., 2020).

Examining the impact on training programs, Meholick et al. found that the OSCORE could be used to identify residents requiring extra surgical simulation training and to assess progress after simulation training (Meholick et al., 2020). Use of the OSCORE impacted standard-setting in an internal medicine OSCE with more residents labelled as failing using the single GRS OSCORE compared to traditional OSCE ratings (Halman et al., 2020). Both of these studies were of moderate methodological quality. In a high methodological quality, national study of emergency medicine competence committee decisions using single GRS OSCORE assessments to guide resident promotion, residents required longer than predicted training time to advance through training, while promotion decisions were based on less assessment data points than recommended. There was also large between-program variability in terms of number of assessments collected and promotion timelines. (Thoma et al., 2020).

Discussion

Having used systematic review methodology to search, identify, appraise and abstract the original research studies, we now articulate the validity argument for the OSCORE, separating out the multi-item OSCORE (either original or modified) implemented in surgical specialties from the single GRS OSCORE implemented in non-surgical specialties.

There is a reasonable validity argument for the multi-item OSCORE in surgical specialties, grounded in an interpretation/use argument of assessing surgical competence as readiness for independent performance for a given procedure. The evidence is predominantly from single-institution studies, across a mix of simulation and clinical contexts, and heavily weighted towards orthopedics and general surgery. The individual studies were of moderate to high methodological quality. The scoring, generalization and extrapolation inferences are well-supported. In terms of implications, there is reasonable data that the OSCORE can feasibly be implemented and effectively used in training programs, and that residents and faculty alike find that it is an acceptable tool. Only one study commented on how the OSCORE was able to identify those in need of more practice at a given procedure (Meholick et al., 2020). It should be noted that in the original study (Gofton et al., 2012), the OSCORE was not intended to be used to make summative decisions about promotion or independent practice, instead focussing on readiness to perform a given procedure

independently. As such, neither the original study nor the additional available evidence supports use of the OSCORE in summative decisions about the promotion of residents through their training program.

Taking a deeper look at some of the issues raised under the scoring, generalization and extrapolation inferences, although the colloquial anchors were intended to encourage raters to use the entire scale (Gofton et al., 2012), there is evidence of range restriction towards the higher end of the scale. Despite limited descriptions and assessment of rater training, generalizability studies did not report major error variance due to raters. From a post-positivist perspective, this may suggest that the construct-alignment of the scales mitigates the need for rater training and/or minimizes the impact of rater bias. (Crossley et al., 2011; Weller et al., 2017).

Alternatively, the unblinded assessment design and rater familiarity with a learner may confound this finding. However, adopting a constructivist lens consistent with the reflexivity of our team, we embrace variability between raters (i.e., we expect that raters, based on different experiences and expertise, would hold different views of a resident's performance). From this constructivist perspective, the lack of variability between raters is unexpected. Possible explanations include, but are not limited to, rater training discourages variability in perspectives or the OSCORE does not encourage varied perspectives on performance.

OSCORE assessments completed in the clinical environment require a significant and potentially prohibitively larger number of assessments to achieve reliability compared to those completed in simulated settings. For surgical cases, reliability can be achieved with a relatively small number of ratings per resident, making it a potentially effective and time-efficient assessment tool for surgical residency programs.

Although there is reasonable evidence for the extrapolation inference, it should be noted that this is largely in the form of expert-novice differences. Although it is reassuring to know that senior residents have higher OSCOREs than junior residents, there may be many confounding factors as to why this is the case (Cook, 2015). As such, expert-novice differences are necessary but not sufficient to support a strong validity argument under the extrapolation inference. No further studies showing expert-novice comparisons are needed.

The original intent of the OSCORE was to assess overall surgical competence, not simply technical skill (Gofton et al., 2012). However, the OSCORE demonstrated moderate to high correlations with other surgical skills assessments. This brings into question whether all surgical performance-based tools, including the OSCORE, are assessing the same construct of surgical competence.

In contrast to the validity argument supporting the multi-item OSCORE in surgical contexts, limited validity evidence exists for the single-item OSCORE in non-surgical contexts. A clear interpretation/use argument has not been articulated in these contexts, although the underlying assumption seems to be readiness for independent performance. There is limited sampling across specialties, programs and centres. The non-surgical studies raise many issues that require further study. It is unclear if the single GRS OSCORE anchors represent a different construct than traditional behaviour-based anchors. Furthermore, there was a lack of correlation between performance in simulation and clinical contexts (Prudhomme et al., 2020). Interestingly, there is very concerning implications evidence in the non-surgical contexts compared to the surgical studies. Two qualitative studies demonstrated mixed impacts of OSCORE assessments on resident behaviours (Dudek et al., 2019; Martin et al., 2020) and highly variable impacts on decision-making regarding resident promotion across emergency medicine training programs (Thoma et al., 2020).

Limitations

We explicitly excluded the Ottawa Clinic Assessment Tool and the Ontario Bronchoscopy Assessment Tool, which were derived from the OSCORE. These tools include scoring items that deviate significantly from the original OSCORE; they may be assessing different constructs. Our synthesis is also limited by the modest methodological quality of the original studies. Notable factors that negatively affected the quality of studies included a single group cross-sectional or post-test only study design, unblinded raters and limited sampling across institutions. We also included simulation-based studies, although it may be argued that the purpose of the OSCORE is for workplace-based assessment where readiness for independent performance is a clearer construct than in the controlled simulation setting. Finally, although we hold a constructivist stance on workplace-based assessment, the bulk of the research into the OSCORE sits firmly in a post-positivist perspective which limited our data interpretation.

Implications for educational practice and research

Acknowledging that there has been confounding in practice between the original IUA of the OSCORE (i.e., "Can this resident perform this procedure independently?") compared to an entrustment supervision decision regarding the procedure (i.e., "how much supervision did I provide to this resident to perform the procedure independently?"), we believe the available evidence does support the use of the OSCORE for ad hoc (in-the-moment) entrustment decisions of surgical procedures by frontline supervisors. The language of the OSCORE anchors aligns with retrospective entrustment supervision decisions ("I had to do it" through to "I did not need to be there") (Ten Cate et al., 2020) and there is evidence for a relationship between competence, independence and entrustment in surgical supervision (Ji et al., 2019). Given that the IUA for OSCOREs in the simulation context focusses on competence (as opposed to readiness for independent performance of a procedure), programs should consider interpreting performance in the simulation context differently from assessment generated in the clinical context.

There is little evidence to support the use of the OSCORE by surgical programs for summative assessment decisions, such as determining the progress of a remediating resident or to make promotion decisions. In order to determine if the OSCORE actually predicts readiness for independent practice, studies comparing OSCORE performance to results of post-residency qualifying exams and actual performance in independent practice are required. These comparisons will take time to develop. Furthermore, if the OSCORE continues to be used in simulation contexts, more validity evidence is required examining the relationship to authentic clinic performance.

An abundance of caution is required in widespread implementation of the OSCORE into non-surgical contexts, given the very limited available evidence. There is a pressing need to articulate the interpretation/use argument for the OSCORE in these settings, and to determine if the current anchors are construct-aligned to either competence or entrustment supervision or whether they represent a novel construct. Much more evidence is needed under each of the inferences to understand the OSCORE in these contexts. In the Canadian implementation of CBME, the OSCORE has been promoted as a core WBA instrument for assessment of Entrustable Professional Activities (RCPSC, 2021). In this educational landscape, it is important to reflect on the ramifications of our articulated validity argument. While the OSCORE underwent rigorous development in the surgical population in the

original study (Gofton et al., 2012), residency programs should be aware that the OSCORE has yet to be studied in community hospitals and little evidence exists outside of surgical specialties. Validity arguments change across contexts, as validity is not a stand-alone property of the tool, and the argument must be re-examined in the new contexts (Cook et al., 2015). Perhaps most concerning, if competence committees are relying on OSCORE data to make decisions regarding resident progression, the only study in this regard suggests high between-program variability, which could threaten the defensibility of the summative decisions (Thoma et al., 2020). Gathering additional data would inform program-specific standard-setting around best practices for the number of assessments and predicted length of training (Thoma et al., 2020).

Conclusion

This systematic review demonstrates that the OSCORE has reasonable validity evidence to support its use for surgical operative assessment, under the scoring, generalization and extrapolation inferences of Kane's framework. However, a validity argument for the extension to non-surgical contexts is not supported. Evidence to support the implications of this assessment instrument is nascent. We are optimistic that the OSCORE can be an informative and relevant tool for postgraduate learner assessment. However widespread adoption must be informed by concurrent data collection in more diverse settings and specialties.

Acknowledgements We would like to thank Dean Giustini and Vanessa Kitchin who were instrumental in guiding our literature search.

References

Cook, D. A. (2015). Much ado about differences: Why expert-novice comparisons add little to the validity argument. *Advances in Health Sciences Education, 20*(3), 829–834.

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education, 49*(6), 560–575.

Crossley, J., Johnson, G., Booth, J., & Wade, W. (2011). Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. *Medical Education, 45*(6), 560–569.

Cutrer, W. B., Russell, R. G., Davidson, M., & Lomis, K. D. (2020). Assessing medical student performance of Entrustable Professional Activities: A mixed methods comparison of Co-Activity and Supervisory Scales. *Medical Teacher, 42*(3), 325–332.

Dudek, N. L., Papp, S., & Gofton, W. T. (2015). Going paperless? Issues in converting a surgical assessment tool to an electronic version. *Teaching and Learning in Medicine, 27*(3), 274–279.

Dudek, N., Gofton, W., Rekman, J., & McDougall, A. (2019). Faculty and resident perspectives on using entrustment anchors for workplace-based assessment. *Journal of Graduate Medical Education, 11*(3), 287–294.

Fitzpatrick, R., Paterson, N. R., Watterson, J., Seabrook, C., & Roberts, M. (2019). Development and implementation of a mobile version of the OSCORE assessment tool and case log for competency-based assessment in urology residency training: An initial assessment of utilization and acceptance among residents and faculty. *Canadian Urological Association Journal, 13*(2), 4.

Gerull, W., Zihni, A., & Awad, M. (2019). Operative performance outcomes of a simulator-based robotic surgical skills curriculum. *Surgical endoscopy*, 1–6.

Gillis, M. E., Scott, S. A., Richardson, C. G., Oxner, W. M., Gauthier, L., Wilson, D. A., & Glennie, R. A. (2020). Developing and Assessing the Feasibility of Implementing a Surgical Objective Structured Clinical Skills Examination (S-OSCE). *Journal of Surgical Education*.

Gilchrist, T., Hatala, R., & Gingerich, A. (2021). A collective case study of supervision and competence judgments on the inpatient internal medicine ward. *Perspectives on medical education*, *10*(3), 155–162.

Gofton, W. T., Dudek, N. L., Wood, T. J., Balaa, F., & Hamstra, S. J. (2012). The Ottawa surgical competency operating room evaluation (OSCORE): A tool to assess surgical competence. *Academic Medicine, 87*(10), 1401–1407.

Halman, S., Fu, A. Y. N., & Pugh, D. (2020). Entrustment within an objective structured clinical examination (OSCE) progress test: Bridging the gap towards competency-based medical education. *Medical Teacher, 42*(11), 1283–1288.

Hauer, K. E., Boscardin, C., Fulton, T. B., Lucey, C., Oza, S., & Teherani, A. (2015). Using a curricular vision to define entrustable professional activities for medical student assessment. *Journal of general internal medicine*, *30*(9), 1344–1348.

Iobst, W. F., Sherbino, J., Cate, O. T., Richardson, D. L., Dath, D., Swing, S. R., Harris, P., Mungroo, R., Holmboe, E.S., Frank, J.S. & International CBME Collaborators. (2010). Competency-based medical education in postgraduate medical education. *Medical Teacher, 32*(8), 651–656.

Ji, S., Hwang, C., Karmakar, M., Matusko, N., Thompson-Burdine, J., Williams, A.M., ... & Sandhu, G. (2020). Association of intraoperative entrustment with clinical competency amongst general surgery residents. *The American Journal of Surgery*, 219(2), 245–252.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Klasen, J. M., & Lingard, L. A. (2021). The butterfly effect in clinical supervision. *Perspectives on Medical Education*, *10*(3), 145–147.

Lord, J. A., Zuege, D. J., Mackay, M. P., Ordons, A. R. D., & Lockyer, J. (2019). Picking the right tool for the job: A reliability study of 4 assessment tools for Central Venous Catheter Insertion. *Journal of Graduate Medical Education, 11*(4), 422–429.

MacEwan, M. J., Dudek, N. L., Wood, T. J., & Gofton, W. T. (2016). Continued validation of the OSCORE (Ottawa Surgical Competency Operating Room Evaluation): Use in the simulated environment. *Teaching and Learning in Medicine, 28*(1), 72–79.

Martin, L., Sibbald, M., Brandt Vegas, D., Russell, D., & Govaerts, M. (2020). The impact of entrustment assessments on feedback and learning: Trainee perspectives. *Medical Education, 54*(4), 328–336.

Meholick, A. L., Jesneck, J. L., Thanawala, R. M., & Seymour, N. E. (2020). Use of a secure Web-based data management platform to track resident operative performance and program educational quality over time. *Journal of Surgical Education*.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, *18*(2), 5–11.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews, 4*(1), 1–9.

Norcini, J., & Burch, V. (2007). Workplace-based assessment as an educational tool: AMEE Guide No. 31. Medical teacher, 29(9–10), 855–871.

Ode, G. E., Buck, J. S., Wally, M., Scannell, B. P., & Patt, J. C. (2019). Obstacles Affecting the Implementation of the OSCORE for Assessment of Orthopedic Surgical Skills Competency. *Journal of Surgical Education, 76*(3), 881–892.

Prudhomme, N., O'Brien, M., McConnell, M. M., Dudek, N., & Cheung, W. J. (2020). Relationship between ratings of performance in the simulated and workplace environments among emergency medicine residents. *Canadian Journal of Emergency Medicine*, pp 1–8.

Reed, D. A., Cook, D. A., Beckman, T. J., Levine, R. B., Kern, D. E., & Wright, S. M. (2007). Association between funding and quality of published medical education research. *Journal of the American Medical Association, 298*, 1002–1009.

Rekman, J., Gofton, W., Dudek, N., Gofton, T., & Hamstra, S. J. (2016). Entrustability scales: Outlining their usefulness for competency-based clinical assessment. *Academic Medicine, 91*(2), 186–190.

Royal College of Physicians and Surgeons CBD oberservations templates. (2021). Retrieved February 14, 2021, from https://www.royalcollege.ca/rcsite/cbd/assessment/wbas/cbd-assessment-templates-e

Saliken, D., Dudek, N., Wood, T. J., MacEwan, M., & Gofton, W. T. (2019). Comparison of the ottawa surgical competency operating room evaluation (OSCORE) to a single-item performance score. *Teaching and Learning in Medicine, 31*(2), 146–153.

Tavares, W., Kuper, A., Kulasegaram, K., & Whitehead, C. (2020). The compatibility principle: On philosophies in the assessment of clinical competence. *Advances in Health Sciences Education, 25*(4), 1003–1018.

Ten Cate, O. (2020). When I say… entrustability. *Medical Education, 54*(2), 103–104.

Ten Cate, O., Schwartz, A., & Chen, H. C. (2020). Assessing trainees and making entrustment decisions: On the nature and use of entrustment-supervision scales. *Academic Medicine, 95*(11), 1662–1669.

Thanawala, R., Jesneck, J., & Seymour, N. E. (2018). Novel educational information management platform improves the surgical skill evaluation process of surgical residents. *Journal of Surgical Education, 75*(6), e204–e211.

Thanawala, R. M., Jesneck, J. L., & Seymour, N. E. (2019). Education management platform enables delivery and comparison of multiple evaluation types. *Journal of Surgical Education, 76*(6), e209–e216.

Thoma, B., Hall, A. K., Clark, K., Meshkat, N., Cheung, W. J., Desaulniers, P., Ffrench, C., Meiwald, A., Meyers, C., Patocka, C., Beatty, L., & Chan, T. M. (2020). Evaluation of a national competency-based assessment system in emergency medicine: A CanDREAM study. *Journal of Graduate Medical Education, 12*(4), 425–434.

Van Heest, A. E., Agel, J., Ames, S. E., Asghar, F. A., Harrast, J. J., Marsh, J. L., Patt, J. C., Sterling, R. S., & Peabody, T. D. (2019). Resident surgical skills web-based evaluation: a comparison of 2 assessment tools. *JBJS, 101*(5), e18.

Weller, J. M., Castanelli, D. J., Chen, Y., & Jolly, B. (2017). Making robust assessments of specialist trainees' workplace performance. *BJA: British Journal of Anaesthesia*, *118*(2), 207–214.