BMJ Surgery, Interventions, & Health Technologies

Computer-assisted simulated workplace-based assessment in surgery: application of the universal framework of intraoperative performance within a mixed-reality simulation

Philipp Stefan ,[1] Michael Pfandler,[2] Aljoscha Kullmann,[1] Ulrich Eck,[1] Amelie Koch ,[2] Christoph Mehren ,[3,4] Anna von der Heide,[5] Simon Weidert,[5] Julian Fürmetz,[6] Ekkehard Euler,[6] Marc Lazarovici,[7] Nassir Navab,[1] Matthias Weigl [2,8]

To cite: Stefan P, Pfandler M, Kullmann A, *et al*. Computer-assisted simulated workplace-based assessment in surgery: application of the universal framework of intraoperative performance within a mixed-reality simulation. *BMJ Surg Interv Health Technologies* 2023;**5**:e000135. doi:10.1136/bmjsit-2022-000135

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjsit-2022-000135).

PS and MP are joint first authors.

Received 27 February 2022
Accepted 24 August 2022

Check for updates

For numbered affiliations see end of article.

Correspondence to
Philipp Stefan;
philipp.stefan@tum.de

ABSTRACT

Objectives Workplace-based assessment (WBA) is a key requirement of competency-based medical education in postgraduate surgical education. Although simulated workplace-based assessment (SWBA) has been proposed to complement WBA, it is insufficiently adopted in surgical education. In particular, approaches to criterion-referenced and automated assessment of intraoperative surgical competency in contextualized SWBA settings are missing.

Main objectives were (1) application of the universal framework of intraoperative performance and exemplary adaptation to spine surgery (vertebroplasty); (2) development of computer-assisted assessment based on criterion-referenced metrics; and (3) implementation in contextualized, team-based operating room (OR) simulation, and evaluation of validity.

Design Multistage development and assessment study: (1) expert-based definition of performance indicators based on framework's performance domains; (2) development of respective assessment metrics based on preoperative planning and intraoperative performance data; (3) implementation in mixed-reality OR simulation and assessment of surgeons operating in a confederate team. Statistical analyses included internal consistency and interdomain associations, correlations with experience, and technical and non-technical performances.

Setting Surgical simulation center. Full surgical team set-up within mixed-reality OR simulation.

Participants Eleven surgeons were recruited from two teaching hospitals. Eligibility criteria included surgical specialists in orthopedic, trauma, or neurosurgery with prior VP or kyphoplasty experience.

Main outcome measures Computer-assisted assessment of surgeons' intraoperative performance.

Results Performance scores were associated with surgeons' experience, observational assessment (Objective Structured Assessment of Technical Skill) scores and overall pass/fail ratings. Results provide strong evidence for validity of our computer-assisted SWBA approach. Diverse indicators of surgeons' technical and non-technical performances could be quantified and captured.

Conclusions This study is the first to investigate computer-assisted assessment based on a competency framework in authentic, contextualized team-based OR simulation. Our approach discriminates surgical competency across the domains of intraoperative performance. It advances previous automated assessment based on the use of current surgical simulators in decontextualized settings. Our findings inform future use

WHAT IS ALREADY KNOWN ABOUT THIS SUBJECT?

⇒ Computer-assisted assessment in surgical simulation is largely limited to the assessment of psychomotor skills in decontextualized settings without involvement of the operating room (OR) team.

⇒ Conversely, competency-based medical education requires a holistic and authentic assessment of surgical competency in settings mimicking the workplace environment.

WHAT ARE THE NEW FINDINGS?

⇒ The universal framework of intraoperative performance can serve as a seminal guide to development of computer-assisted assessment that encompasses technical and non-technical competencies. Assessment in a simulated mixed-reality workplace environment resulted in high levels of perceived authenticity and yielded strong evidence for validity.

HOW MIGHT THESE RESULTS AFFECT FUTURE RESEARCH OR SURGICAL PRACTICE?

⇒ Computer-assisted assessment of intraoperative competencies in authentic, simulated OR settings complement conventional workplace-based assessment with a safe, controlled and objective assessment approach, without cutting ties with surgical practice.

of computer-assisted multidomain competency assessments of surgeons using SWBA approaches.

INTRODUCTION

With the ubiquitous adoption of competency-based models in surgical education, entrustment decisions for unsupervised practice are based on summative assessment of performance as well as formative assessment to guide residents' acquisition of key competencies.[1 2] Contrary to previous time-based educational models, competency-based medical education (CBME) requires authenic, workplace-based assessment (WBA). WBA shall apply multiple objective measures to guide decisions based on criterion-referenced competency scores with a focus on outcomes within real-world surgical practice[3 4] (see also online supplemental table A1).

Although WBAs are desirable and will be the primary assessment setting for the foreseeable future,[1] they are severely limited by numerous factors including cost, patient safety concerns, or non-standardized settings.[5] Therefore, simulation is proposed as a complementary assessment setting.[6] However, current simulation approaches either rely on subjective methods for assessment[7–9] or solely focus on norm-referenced assessment of technical psychomotor skills (PMS) in decontextualized settings.[10–13] We thus introduce a novel contextualized computer-assisted simulated workplace-based assessment (SWBA) method. Our approach is based on the universal framework of intraoperative performance and enables objective, criterion-referenced competency assessment across the technical—non-technical continuum of intraoperative performance. Framing the assessment in an authentic, simulated work environment, we provide a format that reflects the competency demands of surgical operating room (OR) practice.

The universal framework of intraoperative performance has been proposed by Madani *et al*[14] to define expert intraoperative surgical performance.[14] It draws on a holistic conception of surgical competency, dissolving the contemporary dichotomization of skills as either technical or non-technical. The framework proposes five key performance domains: psychomotor skills (PMS), declarative knowledge (DK), interpersonal skills (IPS), personal resourcefulness (PR), and advanced cognitive skills (ACS). It further emphasizes the key role of ACS in surgeons' intraoperative decision-making and behavior.[14]

Our approach facilitates overcoming current deficits of performance assessment in surgery: observational assessment of surgeons' competence is time-consuming and fraught with bias, as it relies on the judgment of individual assessors. Although subjective judgments can be a valuable source for performance feedback if treated carefully, subjectivity may spur bias. Assessment quality depends predominantly on faculty assessors, their availability, commitment, and training.[1 11 15] For reliable assessment, up to ≥7 expert assessors are required.[16] Furthermore, many contemporary methods result in generic scores that allow only limited insight into performance gaps and do not meet the requirements of individualized, meaningful, and case-specific feedback.[14] Additionally, not all aspects of surgical performance are amenable to visual evaluation by expert observers.[17] To this end, observational assessments are not effectively capturing key aspects of surgical performance that ultimately contribute to procedural outcomes.[17 18]

Current computer-assisted (also termed computer-aided) assessments, on the other hand, are predominantly norm-referenced[19] and focus on psychomotor aspects of surgical technique with insufficient consideration of cognitive and non-technical skills.[11 12] Available assessments are almost exclusively situated within decontextualized, non-OR settings, without any team involvement.[10 13] Lastly, the majority of studies use outdated frameworks to evaluate validity. This can lead to misinterpretations of surgeons' performance and spurious educational decisions.[20]

We therefore adopted the universal framework of intraoperative performance with use of computer-assisted performance metrics for application in a team-based simulated OR setting. Our approach thus contributes to several shortcomings: first, it allows for accurate, objective assessment of surgeons' competence and enables immediate formative feedback.[13] Second, it comprehensively captures competencies within the broad continuum of technical and non-technical performance. Third, it embeds assessment in the larger context of a team-based simulated workplace setting and provides a stimulus format authentically mimicking real surgical practice in the OR. Finally, we establish validity evidence according to contemporary standards,[21–23] which is a prerequisite for systematic implementation of assessment tools in surgical practice.[20]

Objectives

We introduce a novel automated SWBA that is based on an established and current surgical performance framework. Specifically, we aimed to

1. Operationalize the universal framework of intraoperative performance and adapt it to the particular demands of a spine surgery procedure (aim 1).
2. Develop computer-assisted assessment comprising criterion-referenced metrics reflecting these key demands and characteristics of intraoperative competency (aim 2).
3. Implement automated performance assessment in an authentic simulated OR workplace setting involving a multiprofessional team and evaluate validity evidence with particular focus on ACS (aim 3).

METHODS

Design

Our multistage development process and investigation consisted of the three consecutive steps:

Step 1: adaptation of performance framework (pertaining to aim 1, see previous discussion).
Step 2: identification and design of metrics (pertaining to aim 2).
Step 3: implementation in simulated OR for competency assessment (pertaining to aim 3).

We followed the reporting guidelines for healthcare simulation research.[24] This study is part of a larger project on OR workplace simulation in spine surgery.[25]

Procedure

Step 1: adaptation of performance framework

We used the universal framework of intraoperative performance[14] and adapted it to vertebroplasty (VP) interventions. VP is part of postgraduate surgical training curricula[26] and was selected as a frequent, prototypical, minimally invasive spine surgery procedure with clear outcome markers, addressing a broad range of skills and involving the entire multidisciplinary surgical team. The procedure involves percutaneous placement of a large bore, hollow needle through the spinal pedicle and subsequent injection of Polymethylmethacrylate (PMMA) cement to stabilize osteoporotic vertebral compression fractures. We reviewed the framework's competence domains, drawing on an expert-informed cognitive task analysis (CTA),[27] literature,[28–33] and previously published computer-assisted technical skill indicators.[11] Further, we sought to include an adjunct surgical product (SP) outcome score. While not explicitly established in the original framework, it is in line with the authors' proposed 'general aims' of the surgical intervention: 'understanding, removing, or fixing pathology; restoring physiology and function; or restoring or altering anatomy' (Madani A, pp258-9)[14] All key performance indicators relevant to VP were then assigned to the framework's domains by surgical experts. Potential disagreements were discussed until consensus was reached.

Step 2: identification and design of metrics

We then established computer-assisted performance metrics and target thresholds for criterion-referenced performance assessment. This step was based on anatomical characteristics of the actual patient case, published empirical data, and best-practice guidelines. We provided an interactive annotation tool that allowed surgical experts to define case-specific target criteria, for example, optimal areas of entry into the pedicle (online supplemental figure A1) or optimal cement injection sites. All computer-assisted metrics subsequently underwent iterative review and refinement by surgical experts until consensus was achieved.

Building on a previously established and evaluated simulation set-up for image-guided spine surgery,[34] we implemented our computer-assisted assessment into the simulated OR setting. Data available for metrics application comprised preprocedure planning data, intraoperative performance data, expert annotations, and segmentations of patient CT data (ie, labels for relevant anatomy, such as vertebrae, large vessels, spinal cord, facet joints, etc). This allowed direct relation of metrics to surgeons' treatment of patient-specific anatomy.

Step 3: implementation in simulated OR for competency assessment

Setting

A full-scale simulated OR was set up with all necessary equipment. We used a previously evaluated mixed-reality approach to radiation-free, image-guided spine surgery simulation to create an authentic simulated workplace for VP intervention.[34] The simulation environment may also be classified as augmented reality (AR) because, as opposed to virtual reality (VR), it is located rather on the 'real' side of the mixed-reality spectrum, and central components of the OR context are physically represented. Based on CT data of a patient with an osteoporotic vertebral compression fracture, a patient-specific synthetic spine model (L2 and parts of L1 and L3) was three-dimensional-printed with two distinct materials for cortical and cancellous bones.[35] This model was then embedded in synthetic soft tissue (red-coloured gel wax) and covered with synthetic skin (three layers of silicone with different levels of firmness, mimicking skin, subcutaneous fat, and muscle tissue).[34 36] The synthetic patient model was then placed between the two parts of a mannequin phantom in prone position on the operating table and fully draped.

Our set-up included preoperative planning and enabled the subsequent simulation of the complete intraoperative VP workflow—from cut to suture. Our complete surgical team consisted of a surgeon (ie, surveyed participant) and a confederate OR team (ie, team members confederated with the study team, instructed to act on their roles in the simulated procedure) including an anesthesiologist, a scrub nurse, and a circulating nurse (see figure 1). The anesthesiologist's workspace was set up at the patient's head behind the drapes, equipped with vital sign monitors and a ventilator. The C-arm and the X-ray monitor screens were placed opposite to the surgeon and operated by the circulating nurse. Surgeons used a foot pedal to trigger simulated (radiation-free) X-ray acquisition. A mobile instrument table with all necessary instruments was positioned within the scrub nurse's workplace, that is, with marker pen, scalpel, jamshidi needle, hammer, bone cement injector, clamps, and suture material. Additional pictures of our simulated workplace setting are shown in online supplemental figures A3–A5. The tasks and interactions of the team members were, for example, to monitor the patient and report the patient's status to the team (anesthesiologist), to control the C-arm as instructed by the surgeon (circulating nurse), or to hand instruments to the surgeon during the procedure (scrub nurse).

Recruitment procedure and sample

Surgeons were recruited from two university teaching hospitals, following snowball invitations via internal

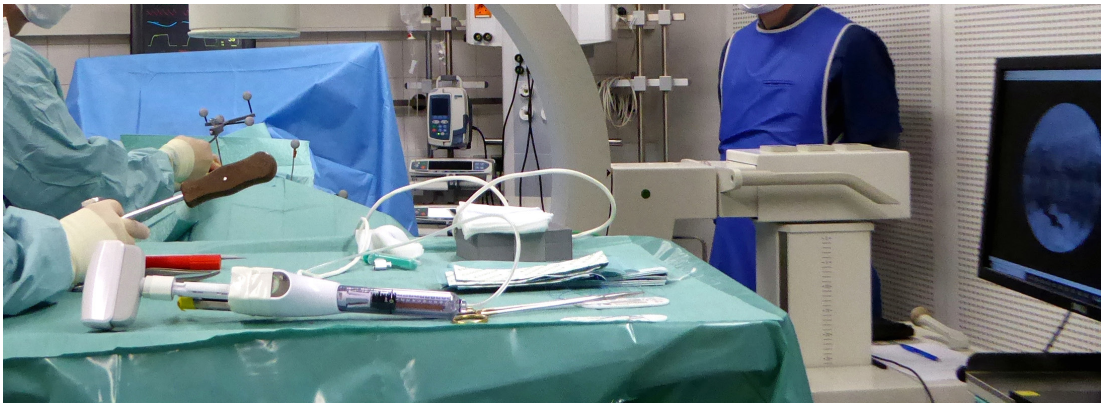

Figure 1 Simulated workplace for vertebroplasty: intraoperative scene during needle insertion with surgical team (anaesthetist behind ether screen).

mailings (ie, we included in all e-mails and information our request to forward the invitation to potentially interested colleagues within the department). The surgeons were incentivized by offering them the opportunity to participate in the assessment and receive feedback on the outcome of the simulated surgical procedure. Eligibility criteria included surgical specialists in orthopedic, trauma, or neurosurgery with prior experience of VP or kyphoplasty interventions.

A total of 11 surgeons participated (72.7% male; six trauma surgeons, three orthopedic surgeons, and two trauma and orthopedic specialists). Their tenure on the job ranged between 0 year and 33 years (mean=7.82, SD=9.38) and between 0 and 200 real VP procedures performed previous to this study (mean=35, SD=9.38).

Simulated procedure

The study was conducted in an academic simulation center. First, surgeons were introduced to the patient case on a planning workstation outside of the simulated OR. Then, they were asked to plan and enter needle trajectories for a bipedicular, percutaneous VP at L2. Afterwards, participants entered the OR and were introduced to the surgical team. All participants were allowed to sufficiently familiarize with the set-up. After the scrub nurse helped participants with gowning, simulation started. It ended with the return of the suture material to the scrub nurse. After the procedure, the participants had the option to receive informal feedback from the study team and additionally review the simulated postoperative CT.

Data collection

1. Procedure planning data: preoperatively, planning data for the VP procedure were obtained from each surgeon, that is, trajectory data for needle insertion (see online supplemental table A2).
2. Intraoperative performance data, serving as input to performance metrics.
   a. Motion tracking data of the needle, the C-arm, and patient model. Data were postprocessed by removing extreme outliers (pairwise distances of subsequent positions 1.5 times the IQR above and below lower and upper quartiles) and applying a Butterworth low-pass filter.[37]
   b. Amount of bone cement injection, measured using a custom developed cement injection device connected via USB.
   c. X-ray/fluoroscopy acquisition, measured by recording foot pedal activation.
   d. Intraoperative teamwork and performance were video-taped (with two opposite cameras).
3. Survey data: After each simulation, surgeons additionally completed a questionnaire evaluation of the simulation environment (rated on a 5-point Likert-scale: 1=strongly disagree, 2=disagree, 3=neither agree nor disagree, 4=agree, and 5=strongly agree).

Measures

1. Standardized observational assessment of surgeons' technical performance was conducted by a chief orthopedic surgeon using Objective Structured Assessment of Technical Skill (OSATS).[25] OSATS is currently 'considered the gold standard in the assessment of technical skills' (Goldenberg and Grantcharov, p125)[38] and consists of (1) a task-specific checklist, (2) a Global Rating Scale, and (3) a pass-fail rating. Assessment of surgeons' non-technical performance was conducted by two trained raters using Observational Teamwork Assessment for Surgery (OTAS).[25] OTAS is a well-established tool for OR teamwork assessment.[39] It measures five non-technical skills essential to surgery. Each skill is measured along behavior exemplars: communication (eg, verbal confirmation of procedure and intra-op requirements), coordination (eg, assessment with nurse on status of instrument preparation before start), cooperation/backup behavior (eg, surgeon responds to questions and requests from nurse), leadership (eg, provides confirmation with nurse for specific surgical requirements ahead of action), and monitoring/situational awareness (eg, reassess set-up and intraoperative requirements in advance). Both OTAS raters underwent observational training before the onset of

the simulation (including pairwise observations with discussion of discrepancies).

2. Computer-assisted metrics were applied to data from surgeons' performances of the full VP procedure with the complete surgical team, except for metrics for DK which were applied to planning data.

Data and statistical analyses

As performance metrics from different categories differ in units of measurement, we normalized values to unit-free scores what facilitates interpretation and score accumulation. Target normalization was applied if target (ideal) and baseline (non-ideal) values were available. Otherwise, we employed ratio normalization, that is, using the minimum and maximum measurement values observed over all participants to normalize single measurements.[40] Cronbach's alpha was computed to analyze intradomain internal consistency, that is, pertaining to what extent scores within one domain measure the same construct. Spearman's correlation was used to investigate interdomain associations of scores and associations of surgeons' experience with observed technical (OSATS) and non-technical (OTAS) performance scores. Wilcoxon rank sum tests were used to test for group differences, for example, between OSATS pass/fail group's associated computer-assisted assessment scores in the performance domains. All statistical analyses were computed with R V.4.0.2.

RESULTS

Step 1: adaptation of performance framework (aim 1)

Relevant indicators were identified and operationalized according to the framework's intraoperative performance domains: PMS, DK, PR, ACS and an adjunct SP score (cf, table 1).

PMS are 'required for neuromuscular coordination of various physical motions and motor functions into meaningful sequences, whose outcome is the accomplishment of simple tasks'(Madani et al, p258).[14] Expert-defined indicators were path lengths to reach relevant anatomy, repositioning attempts, or target accuracy with respect to the preoperative planned trajectory. *DK* pertains to recitable facts, that is, knowledge of relevant anatomy, physiology and pathology, surgical technique and instruments, as well as literature-based knowledge regarding various aspects of patient management such as risk factors and complication rates.[14] We operationalized this domain using surgeons' preoperative planning data (cf, table 1). *PR* describes managing factors with surgeons' ability to impact mental focus, stress, attentiveness, and objectives.[14] Here, we identified total operation time and number of X-rays.[41 42] ACS are defined as advanced mental capabilities such as forward planning, error prevention, and reassessment of the tactical approach with the team. Corresponding indicators therefore quantify compliance and deviation from the expected course of the operation without harmful patient outcome.[14] Indicators relevant to VP include adhering to safety distances, avoiding violation of danger zones or, pertaining to communication, coordination of C-arm adjustment with assistance from the circulating nurse.[27] Pertaining to VP, we defined our SP adjunct as the injection of an adequate amount of cement into the vertebra without causing leakage and avoiding injury of critical anatomical structures.

Step 2: identification and design of metrics (aim 2)

We then identified and designed computer-assisted metrics according to the framework (see previous step 1). Online supplemental table A2 describes resulting metrics' definitions, normalization, and score aggregation. Metrics definitions were based on preoperative planning and intraoperative performance data, that is, surgeons' preoperatively planned trajectories, intraoperatively recorded instrument trajectories, adjustments of the C-arm in spatial relation to patient anatomy, X-ray acquisitions, and cement injection data (cf, online supplemental table A2, score definition). Criterion-referenced targets (ideal performance) and baselines (non-ideal performance) could be established for the majority of metrics (cf, online supplemental table A2). Where possible, we used anatomical properties or limits for target and baseline definition. For example, in the needle insertion step, the target is defined as the maximum anatomically possible distance from the pedicle wall, the baseline defined as zero distance from the wall (see online supplemental table A2, ACS_5: safety margin pedicle wall). For other metrics, we relied on published empirical data and recommendations. For example, the target value of cement injection is defined as the mean of three recommendations for optimal cemented fraction[30–32] multiplied with the case-specific volume of the vertebra (see online supplemental table A2, SP_3: volume of cement deployed in vertebra). Where anatomical limits or recommendations were not available, we established definitions using annotations collected from n=4 experienced and senior surgeons who were very familiar with this procedure. Online supplemental figure A1 shows the resulting labelling of optimal and acceptable areas for pedicle entry, used for metric definition and normalization in the pedicle entry step (see online supplemental table A2, ACS_2: pedicle entry).

Total scores for PMS, DK, PR, ACS, and SP were aggregated by summation of scores, except when highly critical structures (ie, spinal cord, nerve roots, and large vessels) were injured, in which case the SP score was set to 0 (see online supplemental table A2).

Step 3: implementation in simulated OR for competency assessment (aim 3)

We implemented the assessment in a contextualized full-scale OR simulation of a VP procedure with 11 surgeons within a confederate surgical team. We determined the feasibility and sought to gather first evidence for validity of our computer-assisted SWBA.

Mean procedure completion time for the two-sided VP was 27.81 min (SD=9.87, range=9.68–41.87). Overall

Table 1 Universal framework of intraoperative performance five-domain model,[14] our SP adjunct and VP-specific operationalization with corresponding expert-defined performance indicators

Intraoperative performance domain	VP-specific framework operationalization	Related expert-based observational rating
Psychomotor skills (PMS) Visuospatial skills, depth perception, dexterity, bimanual coordination, hand-to-eye coordination	PMS score Path length soft tissue Path length vertebra Needle positioning attempts Accuracy with respect to plan Pedicle entry Vertebra target	OSATS GRS Time and motion Instrument handling
Declarative knowledge (DK) Knowledge of anatomy, physiology, and pathology Knowledge of surgical techniques, procedural steps, and instruments Knowledge of scientific literature	DK score (preoperative planning) Pedicle entry Expert-defined entry area agreement Expert-defined optimal entry point Vertebra target Expert-defined optimal target point Safety margins Distance from pedicle wall Danger zones Pedicle perforation Vertebra perforation Injury of critical structures	OSATS GRS Knowledge of instruments Knowledge of specific procedure
Interpersonal skills (IPS) Teamwork, communication, and cooperation Leadership and management	Not operationalized	OTAS Communication Cooperation/backup behavior Leadership Coordination
Personal resourcefulness (PR) Self-awareness and metacognition Managing modulators of attention, stress, and goals	PR score Time Total X-ray amount	OTAS Monitoring/situational awareness
Advanced cognitive skills (ACS) Surgical planning and error prevention Error recognition, rescue, and recovery	ACS score Intraoperative planning and error prevention X-ray visualization Pedicle entry Expert-defined entry area agreement Expert-defined optimal entry point Vertebra target Expert-defined optimal target point Error recognition, rescue, and recovery Safety margins Distance from pedicle wall Danger zones Pedicle perforation Vertebra perforation	OSATS GRS Respect for tissue Flow of operation OSATS CL Locate entry points Team time-out Skin incisions Guide needle to pedicles Insert needle through pedicles Locate final positions in vertebrae Cement injections Needle removal Skin sutures Final C-arm image control
Surgical product (SP) General aims Understanding, removing, or fixing pathology Restoring physiology and function Restoring or altering anatomy	SP score Cement amount Cement leakage Injury of critical structures	OSATS PF PF rating

Table expanded on Madani *et al* (p258).[14]
CL, checklist; GRS, Global Rating Scale; OSATS, Objective Structured Assessment of Technical Skill; OTAS, Observational Teamwork Assessment for Surgery; PF, pass/fail; VP, vertebroplasty.

procedure success rate was 72.73% (defined as completion without critical patient harm; that is, no highly critical structures were injured and therefore SP score>0).

After each simulation, participants evaluated the simulation via a standardized questionnaire. Participants appraised the authenticity of the simulation and proximity to clinical reality. Surgeons expressed agreement with the overall alignment of the simulation with real VP settings (median=4, IQR=1), strong agreement with workflow replication (median=5, IQR=0.5), and interaction with the OR team members (median=5, IQR=0).

We then determined surgeons' performance and intracorrelation and intercorrelation within and between the framework's performance domains (see table 2). Intradomain internal consistency within performance domains (Cronbach's alpha) was at acceptable levels regarding PMS, PR, and ACS (see table 2). Significant interdomain associations were observed between PMS and PR

Table 2 Performance domain's scores, internal consistencies, and interdomain associations

				Interdomain associations			
	Statistics		Internal consistency	PMS	DK	PR	ACS
Performance domain	M (SD)	Range	Cronbach's alpha	Spearman r (P value)			
Psychomotor skills (PMS)	0.53 (0.17)	0.27–0.78	0.74				
Declarative knowledge (DK)	0.81 (0.05)	0.72–0.87	N/A	0.24 (0.48)			
Personal resourcefulness (PR)	0.34 (0.19)	0.10–0.70	0.66	**0.66 (0.03)**	0.25 (0.47)		
Advanced cognitive skills (ACS)	0.62 (0.14)	0.42–0.80	0.79	0.57 (0.07)	0.01 (0.98)	0.32 (0.34)	
Surgical product (SP)	0.52 (0.39)	0.00–0.90	0.65	0.45 (0.17)	0.10 (0.77)	0.27 (0.43)	**0.85 (<0.01)**

Note: N/A (no variance), n=11 surgeons. Bold if $p<0.05$. Score range 0=worst possible, 1=best possible performance, except for PR and some PMS scores, where this corresponds to worst and best observed performance, respectively (see online supplemental table A1).
M, mean; N/A, non-applicable.

($p=0.03$). Association of domain scores with our adjunct SP was strong for ACS ($p<0.01$; see table 2).

Then we determined associations of automated performance score with experience as well as expert-based observational technical (OSATS) and non-technical (OTAS) assessment scores (see table 3).

Regarding surgeons' experience (job tenure; number of procedures performed), we found strong associations with ACS and with SP. Online supplemental figure A2 shows scores as a function of experience (job tenure) and online supplemental table A3 provides a breakdown of participants into experience levels and their OSATS pass/fail rating, SP and overall score. We further observed medium to strong relationships between expert-rated technical performance (OSATS) and computer-assisted performance assessment: for ACS, SP and overall score. Strong associations between expert-rated non-technical performance (OTAS) and computer-assisted assessment were observed for ACS and SP.

In the next step, we investigated group differences between surgeons pertaining to the OSATS pass/fail rating. Figure 2 shows the corresponding box plot. Differences regarding overall score were statistically significant ($W=0$, $p<0.01$). Per domain differences were statistically significant for ACS ($W=2$, $p=0.02$) and SP ($W=0$, $p<0.01$).

In a final step, we used the OSATS pass/fail rating for standard setting. Using the contrasting groups method,[43] we determined a cut score of 0.54 for our computer-assisted intraoperative competency score (see figure 3). Applying this to our sample, five surgeons were classified as non-competent (below cut score) and six surgeons as competent (above cut score) with no false negatives or false positives in comparison to expert-based OSATS pass/fail judgment.

Lastly, we evaluated validity evidence for our approach according to contemporary, unitary conceptualization of validity[21–23] across the five sources:

1. Content. Representativeness of the achievement domain was established by the rigorous foundation of our simulation on expert-informed CTA[27] and a high degree of contextualization of our simulation approach. This is underpinned by participants' appraisal of the authenticity of the simulation. Test item representativeness is supported by the holistic characteristic of the universal framework of intraoperative performance and synthesis of our operationalization from

Table 3 Relationships between performance domain scores, surgeon's experience, and observational performance ratings

	Association with surgeon's experience		Association with expert-based observational ratings	
	Job tenure	Procedures (n)	Technical performance (OSATS)	Non-technical performance (OTAS)
Performance domain	r (P value)	r (P value)	r (P value)	r (P value)
Psychomotor skills (PMS)	0.10 (0.77)	0.02 (0.96)	0.49 (0.13)	0.15 (0.66)
Declarative knowledge (DK)	−0.14 (0.69)	−0.02 (0.95)	0.34 (0.30)	0.10 (0.78)
Personal resourcefulness (PR)	0.12 (0.73)	0.23 (0.49)	0.57 (0.07)	0.08 (0.82)
Advanced cognitive skills (ACS)	**0.67 (0.02)**	0.49 (0.13)	**0.69 (0.02)**	0.51 (0.11)
Surgical product (SP)	**0.79 (<0.01)**	**0.72 (0.01)**	**0.88 (<0.01)**	**0.79 (<0.01)**
Overall	0.52 (0.11)	0.44 (0.17)	**0.84 (<0.01)**	0.58 (0.06)

Note: n=11 surgeons. Bold if $p<0.05$.
OSATS, Objective Structured Assessment of Technical Skill; OTAS, Observational Teamwork Assessment for Surgery.

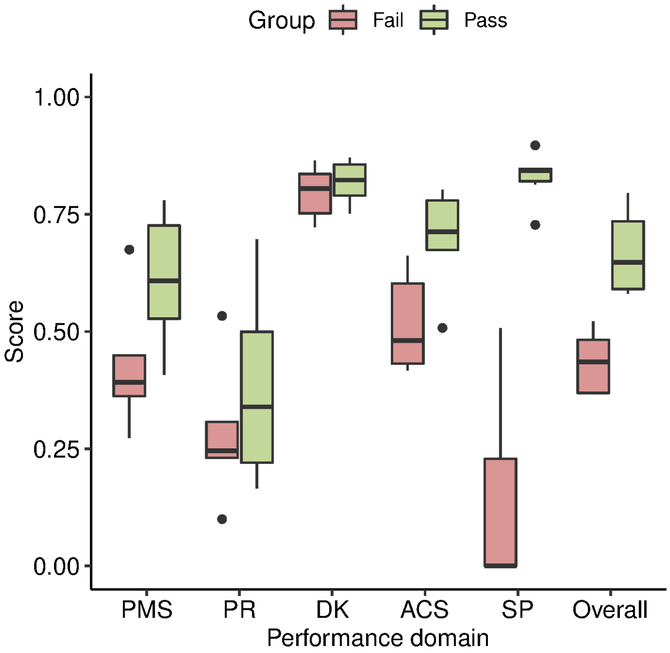

Figure 2 Box plot of surgeons' scores with non-successful (red) and successful (green) performance. Note: attribution to group according to Objective Structured Assessment of Technical Skill pass/fail rating. ACS, advanced cognitive skill; DK, declarative knowledge; PMS, psychomotor skill; PR, personal resourcefulness; SP, surgical product.

multiple sources, that is, CTA, literature and expert knowledge.

2. Response process. Scoring format and combination are informed by the universal framework of intraoperative performance as a profound conceptual foundation. Meaningfulness of score description and interpretation are supported by criterion-referenced scores facilitating interpretation. Our rigorous study design, including standardized and computer-assisted data acquisition, further supports validity of response process.
3. Internal Structure. Our computer-assisted automated approach produced ratings that are reproducible, reliable, and free from bias pertaining to subjective observation. Intraitem/domain internal consistency is at acceptable levels (see table 2). Interitem/domain associations of scores (between performance domains) indicate heterogeneity of constructs across the domains; association with the overall SP outcome was strong for ACS (see table 2).
4. Relationship to other variables. Observed associations of computer-assisted assessment scores with further metrics such as surgeons' experience and technical (OSATS) and non-technical (OTAS) performance were considerable regarding overall score, and strong for ACS and SP in particular (see table 3).
5. Consequences of testing. The educational relevance of our automated assessment approach is substantiated by statistically significant difference in scores between surgeons in the pass/fail groups. Per domain analysis of differences in particular supports the use of ACS and SP scores as discriminators of competent performance for summative decisions. The consequences of using our approach to classify surgeons as non-competent or competent are favourably supported by no false negatives or false positives when opposing computer-assisted classification with expert rater-based judgment.

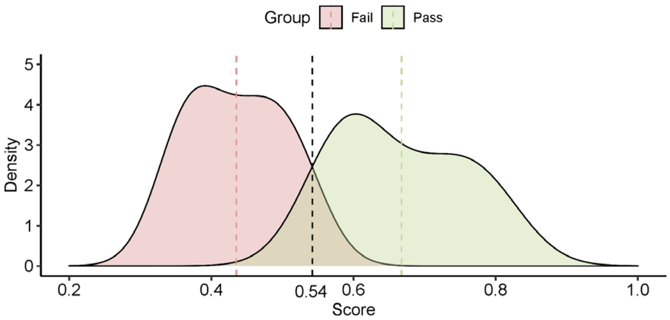

Figure 3 Density plot of surgeon's overall (accumulated) scores with non-successful (red) and successful (green) performance. Cut score determined using contrasting groups method[43] at 0.54 (dashed line). Note: attribution to group according to Objective Structured Assessment of Technical Skill (OSATS) pass/fail rating.

DISCUSSION

Comprehensive, criterion-referenced, and authentic assessment of intraoperative performance is key to evaluate surgeons' competency and quality of surgical care—and a cornerstone of CBME.[1 3 4] We introduce a multistep approach for operationalization of the universal framework of intraoperative performance[14] with definition of respective performance indicators as well as its implementation in surgeons' performance assessment using computer-assisted metrics. Moreover, we demonstrate its feasibility and report first evidence for its validity within automated performance assessment in a simulated procedure. Our findings thus contribute in several aspects to current surgical knowledge base and educational practice.

First, we describe a methodology to use the surgical performance domain constructs proposed through the universal framework of intraoperative performance.[14] After delineating the demands and characteristics of a specific procedure (VP), experts defined meaningful criterion-referenced performance indicators founded on task analysis, anatomical constraints and empirical data. Furthermore, we demonstrate how interactive annotation tools can be used to obtain case-specific definitions based on expert knowledge where such data are unavailable. Our approach therefore can be readily transferred to other surgical procedures beyond the field of minimally invasive spine surgery. Our findings thus serve as a first example for adoption of this widely acknowledged framework and its performance domains to actual surgical interventions and assessments. A wider adoption of the framework would further help to standardize assessments and limit the current inflation of assessment approaches and constructs.[19]

Second, we demonstrate how adoption of the framework can be achieved through automated performance assessments. We developed computer-assisted metrics using preoperative planning and intraoperative performance data in conjunction with procedural and case-specific anatomy characteristics, and experts' annotation data. Our findings advocate that in the era of CBME, computer-assisted assessments represent the next step toward objective analysis of surgical performance that will drastically advance the way surgical trainees learn and are assessed.[44] In addition to objective and reliable scoring, a particular strength of our assessment approach is the direct relation of the scores to procedure-specific and patient case-specific characteristics. This supports meaningful criterion-referenced interpretation of assessment results. Additionally, our approach considers a comprehensive range of competencies within the continuum of technical and non-technical performance. It is thus in contrast to previous computer-assisted assessments in surgery which were almost exclusively limited to norm-referenced assessment of psychomotor aspects of surgical technique.[11 12 19]

Third, we show how computer-assisted authentic assessment of intraoperative surgical competency can be applied in SWBA settings, particularly through highly contextualized OR simulation. To our knowledge, this is the first study to investigate objective, computer-assisted assessment in a simulated workplace that is functionally aligned with a full-scale OR setting. We furthermore included a multiprofessional OR team, what is in contrast to previous decontextualized, single-user benchtop models or VR simulators.[10] This is of particular relevance to CBME with its strong focus on assessments mimicking real surgical tasks conducted in authentic settings 'in the trenches'.[3(p362)] Empirical research has shown that the stimulus format is the paramount factor that determines validity of assessments.[45] For integrated competency assessment, simulations that closely approximate real surgical performance are thus essential.[45] Furthermore, high contextualization minimizes surgeon's distortion of the naturalistic and context sensitive responses, they need to develop for real surgical situations.[46] Our mixed-reality approach allows for a simulated representation of the procedure with little change to the environment and resulted in a functionally active involvement of all OR members. It includes many of the elements identified as central for authenticity of simulated environments,[47] specifically, content drawn from real life (ie, patient-specific simulation using real patient data), interaction and feedback (ie, natural and dynamic interaction with the patient and the team), performance expectations (ie, full performance of the procedure lasting as long as in real life), preparation of the environment (ie, real, functional equipment and devices), presence of a patient manikin, logical and adaptive scenario and sociological fidelity (ie, including all members of the interprofessional team). Accordingly, surveyed surgeons appraised the authenticity of our simulation in general and regarding procedural workflow and interaction with team members in particular.

Fourth, we establish first empirical evidence for validity of assessment based on the universal framework of intraoperative performance.[14] We evaluated our approach in the light of all five sources of evidence for validity.[22 23] Our findings further yield, for the first time, empirical insights into the interactions between the performance domains and their relation to surgeons' experience, competency, and surgical outcome. Interdomain associations of performance scores showed no overly strong relationships. This suggests that, as intended, we measured conceptually different performance types.[14] Notwithstanding, we observed meaningful associations between some domains. This observation warrants further research into potential overlaps and similarities during intraoperative practice, for example, the role of PMS for PR. Regarding surgeons' experience, we identified a substantial association with ACS. Given our limited sample size and sample's intermediate level of experience, however, this finding should be interpreted with caution. Perhaps even more relevant to surgical practice, we found a considerable association between domain scores (PMS, DK, PR, and ACS) and observational technical-skills assessment (OSATS). Post hoc, we assume that the framework's performance scores representing technical aspects of competency are well appraised by expert raters through observational assessments; this association was particularly pronounced for ACS. The consequences of using our assessment approach to classify surgeons' performance as non-competent or competent is favourably supported by no false negatives or false positives in comparison to expert-based OSATS pass/fail judgment in our sample.

Finally, we found that our ACS metric is central to surgical expertise, patient safety, and outcome.[14] We obtained significant group differences for ACS between groups of successful versus unsuccessful performance outcomes (ie, OSATS pass/fail). We also observed moderate yet non-significant associations between ACS and non-technical skills what may tentatively confirm the key role of ACS encoding higher cognitive functions during intraoperative task performance and surgical teamwork.[14] Moreover, high correlations between our SP score and surgeon's experience, technical and non-technical performance support the applicability and validity as a global outcome assessment. Together, these findings suggest that ACS and SP should serve as the central markers of outcome-relevant competency to guide decisions in summative assessment. Immediate assessment outcomes could inform decisions whether a corresponding competency milestone goal has been achieved or if entrustable professional activities (EPAs) can be granted to junior surgeons.[4]

Regarding potential formative assessment in the future, our automated assessment approach facilitates immediate feedback and fulfils the requirement to provide individualized, meaningful, and case-specific guidance.[14] This may help to design curricula 'that target and deliberately train non-experts to think and behave like experts'[14(p263)]; for example, through granular and immediate performance feedback in case of insufficient technical or non-technical performance. A particular benefit emerging from the framework used is that learner feedback can be specifically directed to performance domains and skills requiring special developmental guidance.

Limitations

Our approach has some limitations that should be acknowledged. We laid out our use of performance metrics specifically to minimally invasive spine surgery. Although various of our metrics are generic to surgical performance, such as accuracy, precision, pace, and use of intraoperative imaging, further investigations into further performance metrics are necessary, for example, for open procedures. Although we used a systematic approach to obtain expert consensus on our performance indicators and associated metrics, we cannot infer how specialists in other surgical domains might appraise specific performance outcomes. Given the vast advancements and adoption of technology in surgical performance assessment, we acknowledge that future adoption of computer-assisted assessments will incorporate further performance indicators, for example, intraoperative stress assessed through ambulatory assessments. Our simulation environment was based on a rigorous, in-depth development process to authentically mimic an OR setting as well as multiprofessional surgical practice.[27] Yet, surgeons might not have performed to their fullest potential in the simulation, for example, due to lack of familiarity with the setting or hesitation to being observed. Participants other than the surgeon were confederates to the study team. While this helped to standardize assessment, it also added to the costs and efforts in preparations and simulation planning. Further limitations include our convenience sampling approach and limited sample size. Future investigations should include a larger number of participants and a differentiated analysis of experience levels and subgroups (eg, interns, residents, attendings, etc). Finally, the interpersonal communication (IPC) domain suggested in the original framework[14] was not considered. Future studies need to further apply all performance domains and gather evidence for validity of assessment across different surgical procedures and specialties beyond spine surgery. Particular focus should be devoted to validity evidence in terms of learners' operative performance and patient outcomes in the long term, that is, functional or morbidity outcomes.

Implications for research and surgical practice

Our approach may inform further research in several ways: first, future investigations should scrutinize the utility of the performance domain model in other surgical procedures or conditions (ie, varying patient factors like high acuity, high body mass index, pediatric vs adult, or unusual anatomy). Moreover, our empirical findings concerning the key role of ACS should be corroborated with particular attention to implications for patient safety and surgical outcomes.[14] Second, the range of skills covered should be broadened further toward non-technical aspects. The domain of IPC should be incorporated using computer-assisted assessment, for example, employing machine learning techniques,[48] and association with non-technical observational assessment scores (eg, OTAS) should be investigated. Such assessments need to provide criterion-referenced indicators relating to competencies such as communication, teamwork, or leadership. Moreover, investigations should address how automated assessment results can be best formatted and fed back to support interpretation and provide idiosyncratic guidance, that is, feedback which is 'individualized, meaningful, and case-specific'.[14(p263)] Third, assessments and validity evidence should be further extended to cover the entire surgical team. It should be of particular interest, how intra-team coordination and cooperation can be automatically and objectively assessed and interpreted with regard to procedure and case-specific demands.[14] Assessments therefore need to include team members other than the surgeon as surveyed participants and criterion-referenced performance indicators have to be developed to also capture their intraoperative performance. Fourth, operationalization of the universal framework of intraoperative performance should be implemented in traditional observational work-based assessments in real OR settings, eventually reducing the current inflation of assessment tools[19] for different specialties and aspects of performance (ie, technical and non-technical). Moreover, consistency of participants' intraoperative performance should be investigated in both, the real OR and contextualized simulation setting. For summative assessment, the correlation of assessments in both settings is of particular interest as part of the validity argument.[5] Regarding formative assessments, investigations into individual as well as joint effects of assessment and feedback in contextualized simulation settings on patient outcomes and surgical performance should be of particular interest (ie, by using a prospective pre intervention–post intervention design).

Concerning implications to surgical practice, our approach draws on the principles of CBME and advocates objective criterion-referenced assessment of a comprehensive range of competencies in authentic and contextualized simulated surgical tasks as a complement to traditional workplace based assessments. Our computer-assisted assessment approach may complement rater-based observational assessment, which is time-consuming, inherently subjective, and therefore fraught with different biases.[15] Yet, it shall not be intended as a replacement. Some degree of subjective professional judgment may actually be considered a necessary element of assessment as it can provide valuable feedback and add to a more authentic and holistic appraisal of learners' competence enhancing the validity of assessments.[49]

Computer-assisted assessment in simulated workplaces allows standardizing many of the factors that affect assessments in real workplace environments, for example, effects for raters, the specific patient or scenario, and concurrently enables efficient assessment of key competencies in an OR context.[50] Standardization supports the scoring and generalisation argument; contextualization strengthens the extrapolation argument of assessment score validity.[50] Standardization is of particular interest in formative assessment where alignment with the individual developmental needs of surgical trainees is required. Here, our approach can help to establish required case numbers and cover variation in operative conditions in a structured way.[16] In summative assessment, our approach contributes to establishing a balance between standardization and contextualization of assessment criteria.[50 51]

CONCLUSIONS

We introduce a novel approach for computer-assisted assessment that adopted the universal framework of intraoperative performance in an authentic mixed-reality OR simulation. Our approach meets the fundamentally new requirements of CBME for objective, criterion-referenced assessment of outcome-relevant knowledge application in authentic settings mimicking the real surgical workplace.[1 3 4] While WBA in the OR remains important to reflect complex and variable demands of real-world surgical practice, we showed that our complementary approach of computer-assisted assessment in simulated workplace settings is feasible and perceived as authentic by participants. Our investigation furthermore provides first empirical insights and validity evidence regarding application of the framework's domain model for intraoperative performance assessment. Through the integration of a generic framework with procedure and case-specific adaptability, our approach has the potential to contribute to standardization and contextualization of formative and summative performance assessments in surgical education and practice.

Author affiliations
[1]Chair for Computer Aided Medical Procedures and Augmented Reality, Department of Informatics, Technical University of Munich, München, Germany
[2]Institute and Outpatient Clinic for Occupational, Social, and Environmental Medicine, University Hospital, Ludwig Maximilians University Munich, München, Germany
[3]Spine Center, Schön Klinik München Harlaching, München, Germany
[4]Academic Teaching Hospital and Spine Research Institute, Paracelsus Medical University, Salzburg, Austria
[5]Department of General, Trauma and Reconstructive Surgery, University Hospital, Campus Grosshadern, Ludwig Maximilians University Munich, München, Germany
[6]Department of General, Trauma and Reconstructive Surgery, University Hospital, Campus Innenstadt, Ludwig Maximilians University Munich, München, Germany
[7]Institute for Emergency Medicine and Management in Medicine (INM), University Hospital, Ludwig Maximilians University Munich, München, Germany
[8]Institute for Patient Safety, University of Bonn, Bonn, Germany

Acknowledgements We gratefully acknowledge the support and contributions of Wolfgang Böcker, Nicole Ostermeier, Angelika Richter, Michael Wengler, and Alexander Winkler.

Contributors Study concept and design: PS, MP, ML, and MW. Acquisition, analysis, and interpretation of data: PS, MP, AKu, UE, AKo, CM, AvdH, SW, JF, EE, ML, NN, and MW. Study supervision: PS, MP, and MW. Guarantor: PS.

Funding This work was supported by the German Research Foundation (WE 4682/4-1, NA 620/30-1, and LA 3341/2) and Munich Centre for Health Sciences.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by ethics committee of the Faculty of Medicine, Ludwig-Maximilians-University (approval number 773-15). The participants were informed about the study and its confidentiality measures and they provided written consent prior to data collection.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs
Philipp Stefan http://orcid.org/0000-0003-4386-6456
Amelie Koch http://orcid.org/0000-0002-4389-2868
Christoph Mehren http://orcid.org/0000-0001-9681-6028
Matthias Weigl http://orcid.org/0000-0003-2408-1725

REFERENCES

1 Holmboe ES, Sherbino J, Long DM, *et al*. The role of assessment in competency-based medical education. *Med Teach* 2010;32:676–82.
2 Lindeman B, Sarosi GA. Competency-Based resident education: the United States perspective. *Surgery* 2020;167:777–81.
3 Carraccio C, Wolfsthal SD, Englander R, *et al*. Shifting paradigms: from Flexner to competencies. *Acad Med* 2002;77:361.
4 et alEdgar L, McLean S, Hogan SO. The milestones Guidebook. ACGME, version, 2020. Available: https://www.acgme.org/Portals/0/MilestonesGuidebook.pdf [Accessed 15 Apr 2021].
5 Brydges R, Hatala R, Zendejas B, *et al*. Linking simulation-based educational assessments and patient-related outcomes: a systematic review and meta-analysis. *Acad Med* 2015;90:246.
6 Kneebone R, Nestel D, Yadollahi F, *et al*. Assessing procedural skills in context: exploring the feasibility of an integrated procedural performance instrument (IPPI). *Med Educ* 2006;40:1105–14.
7 Cumin D, Boyd MJ, Webster CS, *et al*. A systematic review of simulation for multidisciplinary team training in operating rooms. *Simul Healthc* 2013;8:171–9.
8 Tan SB, Pena G, Altree M, *et al*. Multidisciplinary team simulation for the operating theatre: a review of the literature. *ANZ J Surg* 2014;84:515–22.
9 Robertson JM, Dias RD, Yule S, *et al*. Operating room team training with simulation: a systematic review. *J Laparoendosc Adv Surg Tech A* 2017;27:475–80.
10 Kneebone RL. Practice, rehearsal, and performance: an approach for simulation-based surgical and procedure training. *JAMA* 2009;302:1336–8.
11 Shackelford S, Bowyer M. Modern metrics for evaluating surgical technical skills. *Curr Surg Rep* 2017;5:24.
12 Hamstra SJ. Workplace-Based Assessment of Procedural Skills. In: Holmboe ES, Durning SJ, Hawkins RE, eds. *Practical Guide to the Evaluation of Clinical Competence. Elsevier Health Sciences*, 2017: 155–64.
13 Vedula SS, Ishii M, Hager GD. Objective assessment of surgical technical skill and competency in the operating room. *Annu Rev Biomed Eng* 2017;19:301–25.
14 Madani A, Vassiliou MC, Watanabe Y, *et al*. What are the principles that guide behaviors in the operating room?: creating a framework to define and measure performance. *Ann Surg* 2017;265:255–67.

15 Driessen E, Scheele F. What is wrong with assessment in postgraduate training? Lessons from clinical practice and educational research. *Med Teach* 2013;35:569–74.
16 Williams RG, George BC, Bohnen JD, *et al*. A proposed blueprint for operative performance training, assessment, and certification. *Ann Surg* 2021;273:701–8.
17 Anderson DD, Long S, Thomas GW, *et al*. Objective structured assessments of technical skills (OSATS) does not assess the quality of the surgical result effectively. *Clin Orthop Relat Res* 2016;474:874–81.
18 Putnam MD, Kinnucan E, Adams JE, *et al*. On orthopedic surgical skill prediction--the limited value of traditional testing. *J Surg Educ* 2015;72:458–70.
19 Szasz P, Louridas M, Harris KA, *et al*. Assessing technical competence in surgical trainees: a systematic review. *Ann Surg* 2015;261:1046–55.
20 Borgersen NJ, Naur TMH, Sørensen SMD, *et al*. Gathering validity evidence for surgical simulation: a systematic review. *Ann Surg* 2018;267:1063.
21 Messick S. Validity. In: *Educational measurement*. 3rd ed. American Council on Education, 1989: 13–103.
22 Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ* 2003;37:830–7.
23 American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). Standards for educational and psychological testing; 2014.
24 Cheng A, Kessler D, Mackinnon R, *et al*. Reporting guidelines for health care simulation research: extensions to the CONSORT and STROBE statements. *Adv Simul* 2016;1:25.
25 Pfandler M, Stefan P, Mehren C, *et al*. Technical and Nontechnical skills in surgery: a simulated operating room environment study. *Spine* 2019;44:E1396.
26 The Royal College of Physicians and Surgeons of Canada. Objectives of training in the specialty of neurosurgery version 1.1, 2014. Available: https://www.royalcollege.ca/rcsite/documents/ibd/neurosurgery_otr_e.pdf [Accessed 15 Apr 2021].
27 Pfandler M, Stefan P, Wucherer P, *et al*. Stepwise development of a simulation environment for operating room teams: the example of vertebroplasty. *Adv Simul* 2018;3:18.
28 Gertzbein SD, Robbins SE. Accuracy of pedicular screw placement in vivo. *Spine* 1990;15:11–14.
29 Cotten A, Boutry N, Cortet B, *et al*. Percutaneous vertebroplasty: state of the art. *Radiographics* 1998;18:311–20.
30 Jin YJ, Yoon SH, Park K-W, *et al*. The volumetric analysis of cement in vertebroplasty: relationship with clinical outcome and complications. *Spine* 2011;36:E761.
31 Nieuwenhuijse MJ, Bollen L, van Erkel AR, *et al*. Optimal intravertebral cement volume in percutaneous vertebroplasty for painful osteoporotic vertebral compression fractures. *Spine* 2012;37:1747–55.
32 Kwon HM, Lee SP, Baek JW, *et al*. Appropriate cement volume in vertebroplasty: a multivariate analysis with short-term follow-up. *Korean J Neurotrauma* 2016;12:128–34.
33 Marchi L, Pimenta L, Oliveira L, *et al*. Distance between great vessels and the lumbar spine: MRI study for anterior longitudinal ligament release through a lateral approach. *J Neurol Surg A Cent Eur Neurosurg* 2017;78:144–53.
34 Stefan P, Habert S, Winkler A, *et al*. A radiation-free mixed-reality training environment and assessment concept for C-arm-based surgery. *Int J Comput Assist Radiol Surg* 2018;13:1335–44.
35 Stefan P, Pfandler M, Lazarovici M, *et al*. Three-Dimensional-Printed computed tomography-based bone models for spine surgery simulation. *Simul Healthc* 2020;15:61–6.
36 Smooth-On. How to make a silicone suture pad. Available: https://www.smooth-on.com/tutorials/creating-silicone-suture-pad/ [Accessed 15 Apr 2021].
37 Yu B, Gabriel D, Noble L, *et al*. Estimate of the optimum cutoff frequency for the Butterworth Low-Pass digital filter. *J Appl Biomech* 1999;15:318–29.
38 Goldenberg MG, Grantcharov TP. The Future of Medical Education: Simulation-Based Assessment in a Competency-by-Design Curriculum. In: Safir O, Sonnadara R, Mironova P, *et al*, eds. *Boot Camp Approach to Surgical Training. Springer International Publishing*, 2018: 123–30.
39 Hull L, Arora S, Kassab E, *et al*. Observational teamwork assessment for surgery: content validation and tool refinement. *J Am Coll Surg* 2011;212:234-243.e1–5.
40 Pollesch NL, Dale VH. Normalization in sustainability assessment: methods and implications. *Ecological Economics* 2016;130:195–208.
41 Hassan I, Weyers P, Maschuw K, *et al*. Negative stress-coping strategies among novices in surgery correlate with poor virtual laparoscopic performance. *Br J Surg* 2006;93:1554–9.
42 Weigl M, Stefan P, Abhari K, *et al*. Intra-Operative disruptions, surgeon's mental workload, and technical performance in a full-scale simulated procedure. *Surg Endosc* 2016;30:559–66.
43 Jørgensen M, Konge L, Subhi Y. Contrasting groups' standard setting for consequences analysis in validity studies: reporting considerations. *Adv Simul* 2018;3:5.
44 Levin M, McKechnie T, Khalid S, *et al*. Automated methods of technical skill assessment in surgery: a systematic review. *J Surg Educ* 2019;76:1629–39.
45 van der Vleuten CPM, Schuwirth LWT, Scheele F, *et al*. The assessment of professional competence: building blocks for theory development. *Best Pract Res Clin Obstet Gynaecol* 2010;24:703–19.
46 Bligh J, Bleakley A. Distributing menus to hungry learners: can learning by simulation become simulation of learning? *Med Teach* 2006;28:606–13.
47 Lavoie P, Deschênes M-F, Nolin R, *et al*. Beyond technology: a scoping review of features that promote fidelity and authenticity in simulation-based health professional education. *Clin Simul Nurs* 2020;42:22–41.
48 Dias RD, Gupta A, Yule SJ. Using machine learning to assess physician competence: a systematic review. *Acad Med* 2019;94:427.
49 Hawkins RE, Welcher CM, Holmboe ES, *et al*. Implementation of competency-based medical education: are we addressing the concerns and challenges? *Med Educ* 2015;49:1086–102.
50 Clauser BE, Margolis MJ, Swanson DB. Issues of Validity and Reliability for Assessments in Medical Education. In: Holmboe ES, Durning SJ, Hawkins RE, eds. *Practical Guide to the Evaluation of Clinical Competence. Elsevier Health Sciences; 2016*, 2017: 22–36.
51 Bates J, Schrewe B, Ellaway RH, *et al*. Embracing standardisation and contextualisation in medical education. *Med Educ* 2019;53:15–24.