BJA

British Journal of Anaesthesia, xxx (xxx): xxx (xxxx)

doi: 10.1016/j.bja.2021.06.038
Advance Access Publication Date: xxx
Review Article

REVIEW ARTICLE

Key dimensions of innovations in workplace-based assessment for postgraduate medical education: a scoping review

Jennifer M. Weller[1,2,*], Ties Coomber[1], Yan Chen[1] and Damian J. Castanelli[3]

[1]Centre for Medical and Health Sciences Education, School of Medicine, University of Auckland, Auckland, New Zealand, [2]Department of Anaesthesia, Auckland City Hospital, Auckland, New Zealand and [3]School of Clinical Sciences at Monash Health, Monash University, Clayton, VIC, Australia

*Corresponding author. E-mail: j.weller@auckland.ac.nz

Abstract

Background: Specialist training bodies continue to devise innovative methods of gathering information on trainee workplace performance to meet the requirements of competency-based medical education. We reviewed recent innovations in workplace-based assessment (WBA) tools to identify strengths, weaknesses, and trade-offs inherent in their design and use.
Methods: In this scoping review, using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines, we systematically searched databases between 2009 and 2019 for WBA tools with novel characteristics not typically seen in traditional WBAs. These included innovations in rating scales, ways of collecting information, technological innovations, ways of triggering WBAs, and approaches to compiling and using information.
Results: We identified 30 innovative WBA tools whose characteristics could be categorised into seven dimensions: frequency of assessment, granularity (unit of performance assessed), coverage of the curriculum, rating method, initiation of the WBA, information use, and incentives. These dimensions had multiple interdependencies and trade-offs, often balancing generating assessment data with available resources. Philosophical stance on assessment also influenced WBA choice, for example prioritising trainee-centred learning (i.e. initiation of WBA and transparency of assessment data), perceptions of assessment and feedback as burdensome or beneficial, and holistic *vs* reductionist views on assessment of performance.
Conclusions: Our synthesis of the literature on innovative WBAs provides a framework for categorising tool characteristics across seven dimensions, systematically teasing apart the considerations in design and use of workplace assessments. It also draws attention to the trade-offs inherent in tool design and selection, and enables a more deliberate consideration of the tool characteristics most appropriate to the local context.

Keywords: competency-based medical education; narrative synthesis; postgraduate medical education; scoping review; workplace-based assessment

Editor's key points

- Competency-based medical education has been adopted by many postgraduate specialist training bodies.
- In this scoping review and literature synthesis, the authors searched systematically for innovations in workplace assessment tools and identified seven key dimensions of these tools: frequency of assessment, granularity (unit of performance assessed), coverage of the curriculum, rating method, initiation of the workplace-based assessment, information use, and incentives.
- The authors provide a framework for categorising tool characteristics and the trade-offs inherent in

Received: 21 April 2021; **Accepted:** 20 June 2021

different choices when designing new methods of workplace-based assessment.
- This framework enables a more deliberate consideration of the tool characteristics most appropriate to the local context.

Postgraduate medical education is moving towards a model of competency-based medical education, in which workplace-based assessment (WBA) plays a central role in learning and in decisions on progression.[1] An early suite of WBA tools was widely adopted by training bodies. These comprise the Mini-Clinical Evaluation Exercise (Mini-CEX), Direct Observation of Procedural Skills (DOPS), multi-source feedback (MSF), and case-based discussion (CbD). However, perhaps because of perceived deficiencies in these traditional WBA tools, training bodies around the world continue to explore innovative methods of generating, aggregating, and reporting information on trainee performance in the workplace.

To explore the range and characteristics of new tools developed to assess trainee workplace competency in postgraduate medical education, we undertook a review of the published literature on recent WBA innovations. In synthesising this literature, we aimed to identify the key characteristics and potential trade-offs inherent in the design of WBA tools and the assessment systems in which they are used.

Methods

We undertook a systematic search and scoping review of the literature using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Following the guidance of Munn and colleagues,[2] we chose a scoping review because we were primarily interested in the qualitative characteristics of WBA innovations, rather than synthesising empirical evidence related to their use. Consistent with the guidelines for the conduct of scoping reviews, quality appraisal of included articles was deemed unnecessary.[2]

Data sources and search constraints

We searched the following databases: MEDLINE, PubMed, PsycINFO, CINAHL, Embase, ERIC, and Scopus. The search was limited to the English language, publication dates between 2009 and 2019 (inclusive), and the human category.

Search terms and inclusion/exclusion criteria

We used the Sample, Phenomenon of Interest, Design, Evaluation, Research type (SPIDER) framework[3] to structure the search terms and the inclusion/exclusion criteria (Table 1). Given the exploratory nature of the search, we did not specify the innovations *a priori* in the search terms, but used general search terms relating to WBA tools, and then manually

Table 1 SPIDER framework for article selection. *'Workplace assessment' OR 'work-based assessment'. †'Mini-Clinical Evaluation Exercise' OR 'Mini-Clinical Evaluation Exercise'/. ‡Direct observation of procedural skills. CbD, case-based discussion; DOPS, Direct Observation of Procedural Skills; EPA, entrustable professional activity; Mini-CEX, Mini-Clinical Evaluation Exercise; MSF, multi-source feedback; PG, postgraduate; SPIDER, Sample, Phenomenon of Interest, Design, Evaluation, Research type; UG, undergraduate; WBA, workplace-based assessment.

SPIDER component	Search terms	Inclusion criteria	Exclusion criteria
Sample	'internship' OR 'residency' OR 'registrar' OR 'junior doctor' OR 'postgraduate medical education'	PG medical specialty trainees; registrar; junior doctor; medical residency; supervisors of PG specialist trainees	UG; non-medical; pre-vocational (foundation years, PG Year 1–2); general practice/family medicine
Phenomenon of interest	'WBA' or variations* OR 'Mini-CEX' variations† OR 'entrustable professional activity/EPA' OR 'DOPS' or variations‡ OR 'direct observation' OR 'multisource feedback/MSF' OR 'clinical encounter' OR 'supervisor report' OR 'peer report' OR 'in-training assessment' OR 'milestone assessment' OR 'portfolio assessment' OR 'field note' OR 'case based discussion'	Innovative ways of collecting information on workplace performance; innovative ways of compiling WBAs for programmatic assessment/high-stakes decisions; innovative alternatives to traditional WBAs (e.g. Mini-CEX, DOPS, CbD, and MSF) and how they are typically combined in assessment portfolios	Simulations; typical use of traditional WBAs (e.g. Mini-CEX, DOPS, CbD, and MSF); minor variation to traditional WBAs; traditional WBAs tailored to specific procedure
Design	Not constrained in search	RCT; cohort comparison; questionnaire/survey; interview/focus group; observational	Conference presentations; abstracts; theses; correspondence; secondary research (e.g. reviews)
Evaluation	Not constrained in search	Comparative or descriptive studies; evaluation or implementation of WBAs; implications for summative decisions; aggregating WBAs; programmatic assessment	Psychometric analyses of common WBAs and their portfolios; technologically intensive precluding general application (e.g. haptic measurement tools)
Research type	Not constrained in search	Quantitative, quantitative, mixed method	None

screened articles for innovations. Our search was not constrained by the design, evaluation, and research type components of the SPIDER, as innovative tools may have been described in a wide range of study types and outcomes of tool evaluation were not the focus of our enquiry. Studies were included if they described a tool for gathering information on trainee workplace performance in the context of specialist medical training and if the tool had innovative characteristics. We used broad inclusion criteria for what qualified as 'innovative'. These were tools other than Mini-CEX, DOPS, CbD, and MSF or their equivalents under different names; WBA tools with novel characteristics (e.g. novel rating scales not typically seen in traditional WBA tools); novel ways of collecting information on workplace performance that were not yet formal tools; novel methods of accessing the tool (e.g. technological innovations); novel ways of scheduling the use of tools across a curriculum (e.g. programmatic assessment); and novel ways of compiling and using information for feedback and assessment (i.e. even if based on traditional WBA tools).

Screening process

Three researchers (JMW, TC, and YC) worked together to screen the abstracts against the inclusion criteria, cross-checking for consistency. One researcher (TC) then screened all included full texts, with each text being read again by at least one other researcher (DJC, JMW, and YC). Inclusion decisions were compared and disagreements were resolved by discussion with a third researcher.

Literature synthesis

Our synthesis was inductive, with no pre-existing model or assumptions. One researcher (TC) summarised the tools according to their key characteristics. The four members of the research team met on two occasions and, through ongoing e-mail exchanges, agreed on these key characteristics. Through discussions with all four researchers, these characteristics were then categorised into agreed dimensions across which the tools and their innovations could be compared. Two researchers (TC and YC) worked together to categorise each tool across these dimensions, whilst two other researchers (JMW and DJC) checked a sample of these categorisations for agreement.

Results

After removing duplicates, we identified 3738 papers. Paper title and abstracts were screened against the inclusion/exclusion criteria, resulting in 193 papers for full-text review. From these 193 papers, 31 papers were included in the review (see PRISMA diagram; Fig. 1). These papers are summarised in Table 2. A glossary of terms is provided in Table 3.

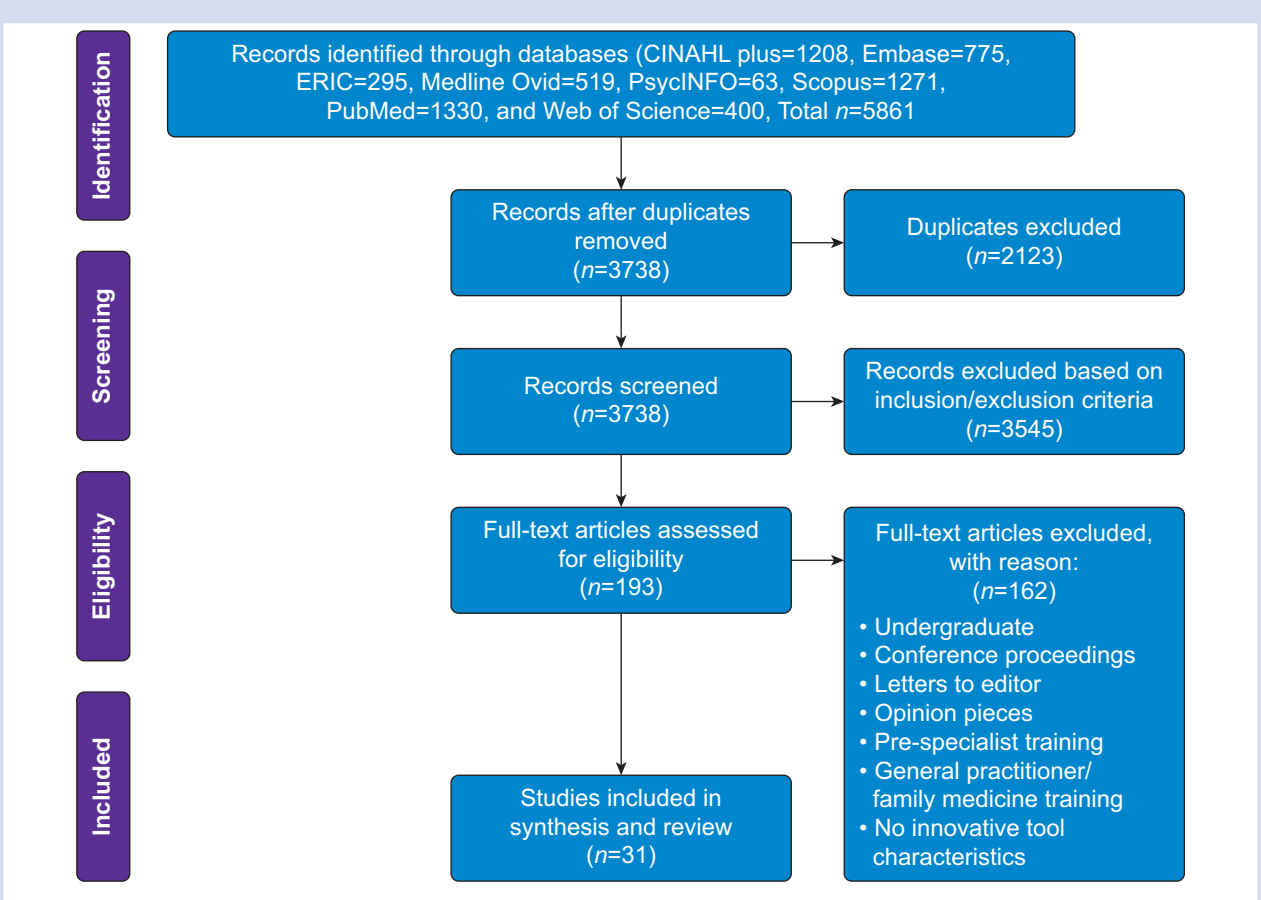

Fig 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram.

Table 2 The final pool of papers, the tools they used, and the outcomes of the studies. CCERR, Completed Clinical Evaluation Report Rating; DEC, daily encounter card; DOPS, Direct Observation of Procedural Skills; ENT, ear, nose, and throat; EOR, end of rotation; EORd, end-of-rotation doctor; EORn, end-of-rotation nurse; EOS, end of shift; EPAsICM, entrustable professional activities of intensive care medicine; ESLE, Extended Supervised Learning Event; IM, internal medicine; JETS, JAG Endoscopy Training System; McMAP, McMaster Modular Assessment Program; Micro-CEX, Micro-Clinical Evaluation Exercise; MSF, multi-source feedback; OCAT, ophthalmology clinic assessment tool; OCAT, Ottawa Clinic Assessment Tool; OCEX, Ophthalmic Clinical Evaluation Exercise; OPA, observable practice activity; OPRS, Operative Performance Rating System; O-SCORE, Ottawa Surgical Competency Operating Room Evaluation; SCO, structured clinical observation; SPR, Surgical Procedure Feedback Rubric; WBA, workplace-based assessment.

Study	Country and responding population	Study design	Tool	Tool description	Outcome
Ehrenfeld and colleagues[4,*]	USA; 60 anaesthesiology residents in one programme	Mixed-method descriptive study; report on how the innovative system was constructed and how the data can be mapped to milestones	Automatic mapping of data from existing perioperative information management system to milestones	Data from all residents' cases automatically organised according to specific process and outcome measures, mapped to milestones, and summarised daily for trainees and programme directors	Tool addresses residents' desire for frequent updates that were perceived previously as insufficient and untimely Can be scaled to varying granularities and coverages, but does not provide data on all milestones Provided clinical outcome data on 24 154 completed anaesthetics across 3 yr
Xu and colleagues[5]	Canada; first- and second-year geriatric medicine trainees at one institution	Mixed-method, descriptive, cohort study; 300 assessments and user-experience surveys analysed for reliability, feasibility, and validity	Consultation letter rating scale	Written communication competencies assessed using six items rated on a 5-point Likert scale, with comments	Quick to complete suggests feasibility, high degree of inter-rater reliability, but raters deemed the 5-point scale not sufficiently discriminatory and trainees valued narrative feedback more than numerical scales
Acai and colleagues[6,*]	Canada; 16 attending physicians from four teaching hospitals under one emergency medicine residency programme	Qualitative, descriptive study; semi-structured interviews studying attending physicians' perceptions of McMAP	McMAP	Programmatic WBA system mapped to CanMEDS roles, comprising high-frequency Micro-CEX and daily global assessments	Programmatic structure perceived to increase the frequency, coverage, and quality of assessments Difficulties giving negative feedback, the possibility of 'gaming' the system, and logistical and technological concerns
Chan and colleagues[7,*]	Canada; pilot group of 15 emergency medicine residents in postgraduate years 1 and 2	Mixed-method cohort study; description of McMAP, quality comparison between 25 randomly selected end-of-rotation reports before and after McMAP, and qualitative focus groups	McMAP	Programmatic WBA system mapped to CanMEDS roles, comprising high-frequency Micro-CEX and daily global assessments	End-of-rotation report quality (assessed using CCERR) increased significantly after McMAP implementation Residents in focus groups note more frequent formative feedback
Chan and colleagues[8]	Canada; 23 second-year emergency medicine residents and 82 supervising physicians (raters) from three teaching hospitals under one training programme	Quantitative cohort study; 1498 global performance score ratings by 82 raters on 23 residents, modelling ratings over time	Global rating score of McMAP	High-frequency assessment of global daily performance using a single, behaviourally anchored rating scale	Residents differ in their starting point (intercept) and rate of progression (slope) Raters introduce substantial variance
Cheung and colleagues[9]	Canada; six experts in resident assessment and eight supervisors	Quantitative; test the extent to which CCERR quality ratings of DECs	DEC	Task-specific items mapped to competencies, rated on a 5-point scale; global	CCERR scores discriminated between the experts' three DEC quality groupings (high, average, and poor)

Continued

Table 2 Continued

Study	Country and responding population	Study design	Tool	Tool description	Outcome
	analysing 60 DECs from emergency medicine	can discriminate between experts' DEC quality groupings		performance item and comments	A generalisability study found most variance (56%) attributable to DEC scores, not to raters
Regan and colleagues[10,*]	USA; 48 emergency medicine residents at one institution	Quantitative cohort study; 5234 assessments collected over 24 months to explore correlations between assessment types	(i) End-of-shift assessment (ii) Electronic end-of-rotation assessment	(i) Rating of shift performance using 9–11 milestone questions from 15 sub-competencies (ii) Two-week electronic global rating of proficiency on 16 sub-competencies (for supervisory doctors: EORd) or four sub-competencies (for nurses: EORn)	EOS and EOR scores were likely to be correlated if they were taken from the same year; EOS scores more strongly correlated with EORd than EORn scores; proficiency of Clinical Competency Committees correlated more strongly with EORd than EOS scores
Braund and colleagues[11]	Canada; nine residents and six faculty in the ophthalmology department of a teaching hospital	Qualitative case study; written feedback and focus groups exploring user experience of four new tools	(i) Ophthalmology field note (ii) OCAT (iii) OCEX (iv) Ophthalmology emergency eye clinic encounter card	(i) Field note: narrative feedback with concern flags, adaptable to encounter or global performance (ii) OCAT: 5-point entrustment scale for six aspects of performance across one half-day clinic, with concern flags and comments (iii) OCEX: fine-grained Likert-scale items measuring performance in one patient encounter, with concern flags and comments (iv) Encounter card: fine-grained checklist items measuring performance in one patient encounter, with concern flags and comments	Field note and OCAT preferred for their simplicity, but concerns around feasibility of high-frequency completion Residents valued narrative feedback over numerical scales, and both residents and faculty valued oral over written feedback despite recognising importance of keeping records Have since implemented a web-based hybrid field note/OCAT
Emke and colleagues[12]	USA; 12 paediatric critical care medicine fellows and 15 faculty supervisors at one unit	Mixed-method cohort study; report on tool construction and utility and validity evidence based on 171 assessments	EPA–OPA tool for paediatric critical care medicine	Twenty sub-EPAs, each with more granular OPAs rated on a 5-point entrustment scale, with comments	Showed adequate utility and validity evidence Narrative justifications for entrustment level consistent with literature on factors influencing entrustment decisions
Toprak and colleagues[13]	Canada; 41 surgical residents and 39 faculty	Quantitative cohort study; description of tool construction, 620	SPR	Twice-weekly rubric-based assessment of performance in a single operation, with items	SPR showed sensible factors in relation to CanMEDS competencies and showed construct validity and

Continued

Table 2 Continued

Study	Country and responding population	Study design	Tool	Tool description	Outcome
	in two training programmes	assessments used to analyse psychometric properties		mapped to CanMEDS roles and rated using behavioural anchors across three levels	the ability to discriminate between levels of training
Hicks and colleagues[14,*]	USA; 165 interns across 15 paediatrics residency programmes in Study 1; 292 interns across 11 paediatrics residency programmes (four from Study 1)	Mixed-method cohort study; report on tool development based on feedback and performance on 873 MSFs and 500 SCOs across one rotation; pilot study on van der Vleuten's[37] utility based on 1241 MSFs and 426 SCOs across one rotation	Web-based system integrating two tools: (i) MSF (ii) SCO	Web-based system that automatically calculates within-competency scores and provides monthly feedback reports based on two tools: (i) MSF: performance across at least 2 days, rated by inter-professional team members on a range of competency-specific items, with comments and global entrustment (ii) SCO: a single observing of learner performance, rated by supervisor on a range of competency-specific items, with comments and global entrustment	Acceptable reliability with four to six instruments Interns reported high-quality, -quantity, -frequency, -specificity, and -timeliness of feedback, and perceived narrative comments as most valuable Concern around faculty buy-in
Anderson[15]	UK; trainee and trainer endoscopists and their training centres	Qualitative, descriptive; description of the development of a web-based endoscopy training and accreditation system	JETS: web-based logbook of DOPS forms	DOPS items mapped to endoscopy competencies; logbook tracks formative period until thresholds are met, then summative period of four DOPS in a month	JETS established and certification and quality assurance processes in place
Donato and colleagues[16]	USA; 80 faculty observers and 73 residents from an IM residency programme at one institution	Retrospective cohort analysis for validity evidence, using 3715 Minicard observations from 2005 to 2011	Minicard direct observation tool	Mini-CEX-style tool assessing more specific behaviours across three competency domains	Validity evidence suggests the tool is apt to identify struggling residents; action-oriented feedback present in 50% of the completed tools suggests useful as formative tool
Anderson and colleagues[17]	USA; 92 surgeons and 150 residents submitted from seven training programmes	Mixed-method cohort study; description of WBA system improvement, drawing on performance data from 3880 assessments and user feedback over a 3 yr period	Two tools integrated in a web-based platform: (i) EPA for patient visits (ii) OPRS for operative performance applied within a briefing, intraoperative teaching and	(i) EPA items describing key elements of a given daily activity, rated using an entrustment scale (ii) OPRS items describing key elements of a procedure with room for narrative comments, completed by both resident and faculty	Observed increase in frequency, quality, and timeliness of assessments, and increase in perceived acceptability of EPAs across the implementation period

Continued

Table 2 Continued

Study	Country and responding population	Study design	Tool	Tool description	Outcome
			debriefing model		
Cooney and colleagues[18,*]	USA; surgical residents in plastics and reconstructive surgery	Mixed-method pilot study proposal; compare traditionally trained residents with trainees under redesigned system	(i) Next Accreditation System milestone assessments (ii) Operative entrustability assessment	(i) Milestone assessments: monthly ratings across multiple sub-competencies, rated on 5-point behaviourally anchored scale from novice to expert (ii) Operative entrustability assessment: global entrustment rating by all faculty surgeons for all witnessed procedures performed by residents	NA: proposal to implement this innovation Intention to create large quantity of data that cross-reference between operative entrustability and milestones
Townend and colleagues[19]	UK; 701 trainees and 750 assessors in emergency medicine	Quantitative cohort study; 1390 assessments analysed for reliability	ESLE	Non-technical skills assessed across 12 domains rated with 4-point scale and comments, within a 3 h observation and feedback episode	Most variation attributable to trainee's ability; G-coefficient of 0.80 with three ESLEs by two or more assessors
Kumar and colleagues[20,†]; Danino and colleagues[21]	UK; 10 ENT surgical trainees across five ENT departments; N trainers not noted	Qualitative, descriptive pilot study; survey and informal discussions assessing user experience	Ward Round Assessment Tool	Records characteristics of the round and rates key components on a 3-point scale with room for comments	Trainees felt it promoted teaching and improved ward round performance, but concern was expressed over the ability for consultants to observe full rounds
Fitzpatrick and colleagues[22]	Canada; 17 urology residents and 12 faculty at the urology division of one university	Mixed-method cohort study; report on tool development and survey study on user perceptions	Mobile Ottawa Surgical Competency Operating Room Evaluation	Algorithmically selected cases evaluated on a nine-item surgical evaluation tool using entrustment ratings, with comments	Residents preferred the ease of access of the mobile version, tended to value written comments most Faculty felt it accurately reflected overall surgical skill and had positive impact on training
Smit and colleagues[23]	The Netherlands; 37 staff and 112 residents from eight paediatric programmes	Mixed-method cohort study; description of new system, initial descriptive statistics, and survey of user experience	Evaluation and assessment of residents by supervisors	Five or more staff assigned to each resident per rotation; each completed an end-of-rotation assessment for use in determining entrustment levels on EPAs when thresholds met	High level of user satisfaction and improvement in feedback quality and timeliness, but noted an administrative burden and the challenge of applying disparate information to a yes/no progression in entrustment level for a given EPA
Warm and colleagues[24,*]	USA; 189 IM residents	Quantitative cohort study; descriptive statistics	OPA-based assessment system	Hundreds of OPAs spread across rotations, rated using an entrustment scale with comments and automatically mapped to IM milestones to track progress. Each assessment	Entrustment increased with stage of training; peer and allied health professionals rated trainees higher than did attending physicians; individual residents did not progress uniformly over time,

Continued

Table 2 Continued

Study	Country and responding population	Study design	Tool	Tool description	Outcome
				consists of 8–10 content OPAs and 8–10 process OPAs	suggesting the need for plentiful data points
van Bockel and colleagues[25]	The Netherlands; intensive care trainers and trainees	Qualitative and descriptive; description of tool and its development	EPAsICM	15 EPAsICM informed by CoBaTrICE and CanMEDS competencies, rated on an entrustment scale with reminders of important entrustment requirements	A measure of overall clinical competence as a supplement to existing assessments of individual competencies; this new WBA tool formalises entrustment decisions
Yuan and colleagues[26,*]	USA; 18 nephrology fellowship trainees	Quantitative cohort study; 8257 charts audited over 5 yr	Outpatient encounter chart audits	Chart audit tool assessing milestones within the six competencies in the context of the EPA 'managing the general and transplantation outpatient clinic'; yes/no items with comments	Deficiencies decline with training year and are negatively associated with examination score percentile; thresholds can be used to establish milestone levels and detect underperformers
Park and colleagues[27]	USA; 116 IM faculty, 59 fellows, and 131 peer residents as raters for 142 residents' end-of-rotation evaluations	Quantitative cohort study; generalisability analysis was conducted on 2701 end-of-rotation evaluations measuring 21 out of 22 IM milestones for 142 residents	IM end-of-rotation evaluations	End-of-rotation global performance on 21 of 22 milestones, rated using a 7-point milestone-level scale, with comments	End-of-rotation milestone assessments are good indicators of their respective core competencies and show good reliability with 10 or more observations; fellow and peer ratings may be particularly useful for professional and interpersonal/communication skills
Warrington and colleagues[28]	USA; 324 emergency medicine educators	Quantitative cohort study; descriptive statistics and inter-rater reliability on 324 forms	End-of-shift evaluation forms (also known as daily encounter cards)	Eight forms providing 76 yes/no data points related to milestones within 16 sub-competencies, with comments	Stimulate feedback and are quick to complete, but only slight-to-fair inter-rater reliability, so caution against using as summative assessment
Rekman and colleagues[29]	Canada; 44 staff surgeons assessing 79 residents at the level of 'generalist' surgeon	Mixed-method cohort study; generalisability analysis; qualitative analysis on feedback from 132 completed assessments	OCAT	Nine to ten items reflecting global performance across a full clinic, rated using an entrustability scale, with comments	OCAT perceived to be useful and to promote formative feedback; most variance in scores attributable to resident performance, and only three observations needed per trainee for good reliability
Hanson and colleagues[30]	USA; faculty and trainees in the Department of Paediatrics at one university	Review article with description of proposed system and pilot period	Narrative evaluation system	Narrative descriptions of trainee performance at the procedure and EPA level, thematically analysed by expert faculty then mapped to ACGME milestones	Narrative comments provide meaningful clinical data about both strengths and weaknesses that richly describe competence and areas for improvement Still work to be done on managing the amount of qualitative data, building a culture of feedback, addressing concerns around acceptability compared with rating scales
Turner and colleagues[31,*]		Quantitative cohort study; descriptive statistics and	Paediatric milestone ratings	End-of-rotation rating of developmental level on each	Faculty assessors judged interns to be at a higher developmental level

Continued

Table 2 Continued

Study	Country and responding population	Study design	Tool	Tool description	Outcome
	USA; 179 interns and 32 sub-interns from 17 paediatrics programmes	comparison between interns and sub-interns		of the sub-competencies, based on electronic compilation of MSF and SCO tools	than sub-interns in most sub-competencies, supporting the validity of the scale
Kameoka and colleagues[32]	Japan; 13 senior residents from three teaching hospitals; five reviewers	Quantitative, pilot comparative study; comparing new tool to programme director evaluation to evaluate criterion validity	Peer review of performance based on residents' completed charts	Visiting interns evaluate residents' charts using a 15-item form that rates patient care process and outcome on a 5-point scale	Correlation coefficients suggest good criterion validity for clinical reasoning and history taking, but not for physical examination and attitude towards patient
Van Heest and colleagues[33]	USA; 294 residents and 370 faculty from 16 orthopaedic surgery programmes	Quantitative cohort comparison study; comparing 1150 O-SCORE assessments with 1186 P-score evaluations and assessing user experience	(i) O-SCORE assessments (ii) P-score evaluations	(i) O-SCORE assesses eight domains of surgical performance on a 5-point entrustment scale, nine-question formative evaluation (ii) P-score evaluates global surgical competence for a case on a single 5-point item	Faculty and residents valued the tools, with residents preferring the P-score and faculty preferring the O-SCORE. Both P-score and O-SCORE discriminated between levels of training and have since been combined into one tool: the 'OP score'
Weller and colleagues[34]	Australia and New Zealand; anaesthesia trainees	Quantitative cohort study; 7808 assessments analysed for reliability	Mini-CEX with entrustment scale	Global performance for a case rated with a 9-point entrustment scale, with comments on domains of practice and overall performance	Entrustment increased with duration and stage of training; moderate reliability with feasible number of assessments; adjusting for expected entrustment improved reliability; detects underperforming trainees
Weller and colleagues[35]	Australia and New Zealand; 80 anaesthesia trainees and 84 assessors across three teaching hospitals	Quantitative cohort study; 338 assessments analysed for reliability	Mini-CEX with entrustment scale	Ten domains of practice rated with a 9-point entrustment scale, and overall performance rated according to entrustment and against expected level of performance	Entrustment ratings significantly improve mini-CEX reliability over ratings against expected level of performance; correcting for expected entrustment given case difficulty improved reliability further

Utilised big data sets. *Kumar and colleagues[20] appear to be the full-text follow-up to the conference proceedings in Danino and colleagues.[21] They appear to be based on the same study, use the same tool, and involve the same authors (albeit in a different order). Given that the tool presented in Danino and colleagues met all other inclusion criteria, the Kumar and colleagues[20] reference was added to supplement the conference proceedings by Danino and colleagues.[21] that otherwise would have been excluded.

Table 3 Glossary of workplace-based assessment terms.

Acronym	Tool
OPA	Observable practice activity
EPA	Entrustable professional activity
OPRS	Operative Performance Rating System
EPA–OPA	Observable practice activity nested within an entrustable professional activity
McMAP	McMaster Modular Assessment Program
JETS	JAG Endoscopy Training System

Literature synthesis

We identified 30 innovative WBA tools and seven dimensions against which we could categorise the tool characteristics and their use. These dimensions were frequency, granularity, coverage, rating method, initiation, information use, and incentives (see Table 4). In the following section, we define and explore each of these seven dimensions. The summary characteristics of the 30 tools are available in Supplementary Appendix 1.

Frequency

We defined *frequency* as the number of times an assessment tool is used or intended to be used, as prescribed or encouraged by a given training programme or in a given study. The 30 tools reviewed here range in assessment frequency from every patient encounter[4,5] to daily assessments,[6–10] twice weekly,[11–13] weekly,[14,15] fortnightly,[16] monthly,[17,18] every 3–6 months,[19] and yearly.[20,21] Some tools use a customisable algorithm that selects for assessment every *n*th case for common procedures and every case for rare procedures.[22] Others use an accumulation of *ad hoc* formative assessments until performance standards are met, followed by an end-of-rotation assessment[23] or a period of summative assessments at a prescribed frequency.[15]

The reported advantages of high-frequency assessments included the ability to document and model progress over time,[6,7,9,15,24] the timely flagging of performance concerns,[7] and the increased opportunity to cover or reassess a wide variety of tasks and competencies.[6,7,18] However, a reported disadvantage of high-frequency assessment systems was the difficulty in collecting, handling, and interpreting large amounts of data.[6–8,12] Further, compared to end-of-rotation assessments, high-frequency end-of-shift assessments had lower correlation with overall performance as judged by clinical competence committees.[10]

Table 4 Key dimensions of workplace-based assessment tools.

Key dimension	Definition
Frequency	Number of times an assessment tool is used or intended to be used over a given time
Granularity	Unit of performance or competence that a tool assesses
Coverage	Breadth of information included in a tool
Rating method	Format in which the tool captures performance information
Initiation	How the assessment starts, and how the assessor and task details are decided
Incentives	Factors that encourage the proper use of the tool
Information use	What happens to the information after it has been recorded

Granularity

We defined *granularity* as the unit of performance or competence that a tool assesses. It can also be thought of as resolution of a tool. Tool granularity ranges from broad aspects of performance and whole procedures down to discrete technical skills. Examples include entrustable professional activities (EPAs),[17,23,25,26] which assess an entire domain of practice, whilst sub-EPAs or observable practice activities (OPAs)[12,24] focus down on single tasks.[12] A tool often contains nested granularities. For example, the items might refer to technical skills that are clustered around competencies or sub-competencies,[9] in which case the tool assesses the coarser granularity of competencies by measuring the finer granularity of technical skills. Granularity also has a temporal element. For example, a tool may measure a snapshot of performance in a single observation (fine temporal granularity),[16] or it may measure global performance at the end of a shift or rotation (coarser temporal granularity)[10,27,28] and across a full clinic.[29]

These coarser granularities are overarching measures of progression towards independent practice. Finer-grained assessments provide a detailed picture of performance, but also increase the amount of data to collect, handle, and interpret. Coarser-grained assessments have lower data demands, but they do not explicitly track the finer-grained justifications for a given rating. Some researchers report that if an overarching assessment is too broad or abstract, assessors can find it challenging to link to observed behaviour.[12,24] For example, the internal medicine EPA 'manage care of patients with acute common diseases across multiple care settings' requires ratings over multiple observations and contexts.[24] Innovations in this dimension therefore also attempt to map finer-grained observable units of behaviour to these broader grained desired outcomes.[4,6,7,9,10,12–16,24,30,31] This may be through inclusion of an explicitly coarse-grained item of overall measure of performance along with the observable units of performance.

Coverage

Coverage refers to the breadth of information included in a tool. If granularity is the resolution (the level of detail included), then coverage is the zoom (how much of the picture is captured). Tools can range from covering only a single component of the training programme (e.g. a specific procedure[20]) to informing across a wide range of competencies at once.[27,32] The clearest picture of overall trainee competence would require both wide coverage and high granularity. However, it would also require large amounts of data and the demands involved in collecting it. Assessment systems that aim for both typically use frequent, brief assessments.[6–8] An innovation in this dimension involves a wide-coverage high-granularity approach using learning analytics to systematically map assessments to all key performance measures across the training programme.[6–8]

Rating method

Rating method is the format in which the tool captures performance information at its specified coverage and granularity.

The three types of rating method in the tools reviewed here are checklists, Likert-type scales, and narrative comments. Checklists typically rate whether something was done or not done.[20] Likert-type scales are usually 5- or 7-point ordinal scales that quantify a characteristic, such as proficiency,[10] level of expertise,[33] or entrustment.[34] Narrative comments are often included in a tool to justify the quantitative ratings or to provide constructive feedback.

Innovations in rating methods include milestone levels[14] and entrustment scales.[34] Milestones specify expected trainee progression across all components of practice. Milestones are scored on a Likert-type scale from 1 to 5, and each score for each component of practice has a description of behaviour representing that level. The description for Level 1 represents behaviour expected of a new trainee, and the description for Level 5 represents behaviour expected of a specialist practitioner. The graduation target is Level 4 across all milestones. These behavioural anchors help to translate the observed performance into the format of the scale. Milestone scales, by definition, compare trainee performance to expected performance, so they are useful for identifying which trainees are on track and which require extra assistance.

Entrustment scales rate the extent to which a trainee is deemed safe to independently perform a given task or procedure. They range from the trainee just observing the task, to the trainee being entrusted to perform the task independently with distant supervision, and finally to the trainee being able to supervise others doing the task. Entrustment scales are similar to milestones if the expected levels of entrustment of particular tasks or cases are specified across levels of training.

Entrustment scales vary in their language and can refer to independence, autonomy, or level of supervision required. They can also refer to retrospective entrustment (how much supervision the trainee needed for that procedure) or prospective entrustment (how much supervision the trainee needs for the next similar procedure). Entrustment scales have been shown to improve the reliability of a tool,[34,35] and may align more closely with how supervisors implicitly make judgements on trainee competence.[12]

Innovations also exist in narrative rating methods. Narrative comments can provide justification or context for the scores on the rating scales, but can also be central to the assessment tool. For example, the field note of Braund and colleagues[11] offers mainly narrative feedback with just one global prospective entrustment rating for the next similar procedure. Hanson and collegues[30] take the centrality of narrative comments even further by arguing that supervisors should only record rich narrative assessments of trainee performance. Experts would then thematically analyse these descriptions to map performance to broader markers of progression, such as milestones and competencies. Narrative comments are usually highly valued by trainees for their detailed feedback.[11,14,22,30,36] They also provide flexibility in the content that supervisors record, because they are not bound by items or scale.[30] However, the large amount of narrative comments required poses feasibility issues for data collection and analysis.[30]

A final innovation in the method of recording trainee workplace performance is in mapping real patient data to broader markers of competence. One assessment system takes data already collected by the perioperative information management system, runs it through a code that compares it to specified thresholds, and then maps these clinical data to relevant trainee milestones.[4] Although this system does not inform all milestones, it does inform some for very little ongoing administrative effort.

Initiation

Initiation refers to how the assessment starts and, where relevant, how the assessor and task details are decided, and is generally either the trainee, assessor, or programme.[6,17] Trainee initiation may represent a learner-centred approach, but is also vulnerable to trainees selecting easy tasks and lenient assessors.[6] Assessor initiation overcomes these biases, and in one system[15] the trainees may not be aware they are being assessed. Programme-initiated assessments refer to constraints on assessment selection imposed by the programme structure[6,7] or by the nature of the tool.[9] Programmes can also use customisable algorithms that select the procedures to assess.[22]

In some cases, both trainee and assessor are involved. For example, the Operative Performance Rating System (OPRS) described by Anderson and colleagues[17] allows both trainee and assessor initiation, and both complete the form. Similarly, both trainee and assessor complete the daily operative entrustability rating described by Cooney and colleagues,[18] after the trainee initiates the assessment and completes the case information. Further, the tool described by Emke and colleagues[12] has both trainee and assessor choosing which broad area of practice (EPA) and which sub-task within that (OPA) to assess. In some cases, the trainee and assessor are electronically linked to the same tool, creating a 'dual responsibility' to complete it.[12]

Incentives

Incentives are defined as factors that encourage the proper use of the tool. Innovations in this dimension include financial incentives, ease of access, and tool design. The Minicard direct observation tool described by Donato and colleagues[16] offers a financial bonus for supervisors who complete one assessment per week. Many of the tools included for review are accessed online through mobile technologies. This makes tools more accessible and quicker to complete, and removes the need for subsequent data entry. Finally, tools can contain affordances that promote better use; for example, they may include a prompt to ensure feedback has been discussed with the trainee in real time[9] and to ensure the feedback is action oriented.[16]

Information use

Information use refers to what happens to the information after it has been recorded on the tool. Innovative approaches describe how performance information is fed back to the trainee and their supervisor, and how it is used for summative assessment, and the relationship between these different purposes.

A number of web-based tools offer immediate performance feedback to trainees and their supervisors online.[4,6,7,15,17] The McMaster Modular Assessment Program (McMAP)[6–8] allows trainees to view aggregated information, which, in the context of their high-frequency programmatic assessment system, graphically displays progress over time and in comparison to the rest of the peer group. The McMaster Modular Assessment Program also automatically generates draft narrative reports at the end of each month-long rotation, which are then thematically analysed to produce qualitative end-of-rotation

reports. These qualitative reports inform tailored feedback for the trainees, which can be used to flag and remediate marginal performances. When used for summative assessments, high-frequency data can automatically map trajectories against thresholds or cut points, whilst narrative data require more dialectical, jury-based review (both of which appear in McMAP).

Similarly, the JAG Endoscopy Training System (JETS) creates anonymous online feedback for the trainee upon completion of a formative assessment.[15] Further, supervisors can track individual and aggregated progress, comparing performance to benchmarks and allowing early identification of sub-optimal performers. The web-based EPA and OPRS tools described by Anderson and colleagues[17] offer similar online performance analytics to the trainee and their supervisors. Both JETS and the EPA and OPRS tools contribute to summative assessment. The web-based integration of MSF and structured clinical observation (SCO) tools described by Hicks and colleagues[14] automatically produce a score for each competency based on the item scores in both tools. These are automatically reported back to trainees along with narrative comments every month. A summary of these monthly reports is ultimately forwarded to the Clinical Competency Committee to make progression decisions.

Whilst some tools are designed just for formative assessment,[16] many produce data that are initially used for formative feedback, but ultimately the data also contribute to summative decisions.[14,24] A drawback of this approach is trainee perception that the formative assessments are higher stakes, which may create unnecessary pressure in the context of generating constructive feedback for learning. The JETS[15] resolves this problem by using their tools in a formative way until performance thresholds are met, followed by a number of summative assessments using the same tools across a month.

Discussion

In this scoping review, we identified 30 innovative WBA tools published in 31 studies over the decade to December 2019. We found that the characteristics of the WBAs could be categorised into seven dimensions: frequency, granularity, coverage, rating method, initiation, incentives, and information use. These dimensions have multiple interdependencies and trade-offs. Furthermore, the choice of tools for workplace assessment will depend on one's philosophical stance on assessment.

Interdependencies and trade-offs

Choices in WBA tools can be considered using these dimensions matched to the educational context, available resources, and the intended purpose of the assessment. Making a choice on one WBA dimension may constrain the feasible choices on other dimensions and has consequences for the user experience and the data produced.

For example, high-coverage and high-granularity assessment can produce a broad and detailed picture of trainee competence, but high-frequency assessment may be required to accumulate enough data to achieve this. However, to be acceptable, high-frequency assessments must each be simple and quick to complete, thereby constraining each assessment event either to high coverage and low granularity (e.g. global score for the day), or high granularity with low coverage (e.g. ratings of a very specific aspect of competence for a given encounter). Further, the amount of data produced through high-frequency assessments requires administrative effort to compile and interpret. In addition, the overall assessment system needs to be tightly controlled because of the constraints of high-frequency assessments (i.e. one simple score for the day, or a score for a very discreet task) to ensure all aspects of competence are covered over time.

In another example, a preference for narrative data (on the 'rating method' dimension) also has implications for the other dimensions. Narrative data may be high or low frequency, but high-frequency narrative comments would only be acceptable if brief. A brief comment might refer to global performance (low granularity) or a particular aspect of performance (high granularity). Rich descriptions of the trainee's performance take time both to write and to analyse, so would require a lower frequency. Free-form narrative comments allow the assessor to focus on salient aspects of performance without being constrained by the specific question and rating scale. However, training bodies often seek specific information, in which case responses to set items may be considered necessary alongside narrative comments. The interpretation of narrative data introduces an additional layer of subjectivity or judgement. To ensure that these more subjective decisions are defensible, the data collected may need to go through a jury-based, dialectical process.

As in quantum physics, the observer effect influences designs of workplace assessment,[36] and trainee behaviour may change because they are being observed. Although individual WBAs are generally designated as low stakes in that they contribute to high-stakes decisions rather than determine them,[37] any assessment of observed performance may be considered 'high stakes' by the trainee.[38] This concern may, to some extent, be overcome with the high-frequency tools, where every moment is a data point. At the other end of the spectrum, some designers have clearly demarcated WBAs as either for feedback only,[15] or to contribute to a high-stakes decision.[15]

In the article, 'Twelve tips for programmatic assessment',[39] the authors state that, 'High-stakes decisions must be based on many data points of rich information, that is, resting on broad sampling across contexts, methods and assessors'. Our review provides some options on the choices available for workplace assessment tools and their use, and the implications for these different choices.

Assessment philosophy

Evident in our review are expressions of the assessment philosophy underlying the different choices, for example holistic *vs* reductionist, trainee agency in the training programme (i.e. who owns the learning), and perceptions of assessment as burdensome or beneficial.

Holistic vs reductionist

The examples provided earlier on assessment frequency and granularity represent hypothetical 'extremes' in the collection of rich data on a trainee's competence. One is characterised by a big data, 'objective' emphasis, with almost constant monitoring. The other favours dialectical narrative, where consensus decisions are made through review of multiple perspectives and potentially contradictory information on a trainee. Assessment choice may often be somewhere in-

between. However, it is interesting that these two extremes roughly reflect the apparently diverging philosophies between studies arising in North America (big data, objective emphasis) and those from the UK, Europe, Australia, and New Zealand (dialectical narrative emphasis). Of the 30 included innovations, nine utilised big data (now flagged in Table 2), and of these, seven were from the USA and two were from Canada. The comparative advantages of these different approaches are likely to emerge over time.

Trainee agency

The philosophical stance that all assessments should be of value for learning and beliefs about the trainee's position as the subject of assessment or a participant in assessment will influence choices on the extent to which trainees have agency in the conditions of their assessments.

Assessment *for* learning suggests the assessment is done *with* the trainee rather than *to* the trainee. A practical expression of this is the dimension of initiation. Trainee agency in choice of assessments provides some benefits to the trainee in terms of their sense of control and encourages self-directed learning. However, the included studies describe a full range from complete trainee control to none, with control by the supervisor or even an external algorithmic decision-making tool within the assessment system.

The assessment philosophy is also revealed in attitudes towards transparency of the assessment process. We view transparency as the extent to which the trainee can access or control the assessment process and the collected data, for example, whether trainees know they are being assessed and whether they see their raw assessor scores and comments.

The opposing stances in these decisions on initiation and transparency appear to hang on trust: can trainees in fact be trusted with their learning and assessment, or will they 'game the system', hiding their deficiencies, seeking assessments after the event on cases that went well, or submitting only those assessments that present them in a good light.[40–42] Taken to the logical conclusion, the consequences of these behaviours would be incompetent graduates with potential for harming patients. With clinicians ultimately responsible for their own performance, at what point should we expect personal responsibility for patient safety to begin? If one of the goals of an assessment system is the development of reflective self-directed clinicians capable of using informed self-assessment to manage their own learning,[43] then to what extent could or should this also be a guiding principle in choices on trainee agency in their assessment?

Burden or benefit

A further philosophical stance to consider is the view of assessment as burdensome or beneficial. The high-frequency assessment scenario implies an environment characterised by near-constant monitoring and assessment. Depending on the local learning environment, this may produce a burden on the work atmosphere, where the roles of 'assessor' and 'assessed' are always at play. In this context, trainees may be concerned that revealing an area that needs development will be recorded in the system as a 'low-competence' data point.

Whilst constant surveillance of workplace performance may be interpreted as threatening, inhibiting, or stressful, the flip side is the demonstrated value of frequent feedback. The local learning environment will or should inform assessment choices. For example, where relationships in the learning environment are trusting and feedback is expected, welcomed, and freely given, more assessments may be beneficial, but where trainees lack trust in the good intentions of their supervisors or supervisors view supervision as a chore, frequent assessments may indeed become a burden.

Recent technological advancements may overcome some of the arguments about assessment as a burden: tools and electronic portfolios can be accessed on mobile devices,[44] voice recognition and automated transcription can relieve the burden of data entry, learning analytics and artificial intelligence techniques can process and interpret narrative data using machine learning, and automatic capture of patient outcome data would bypass the need for additional data entry.[4,45]

Limitations and future directions

This review was limited to English language. We acknowledge potential bias on our inclusion criteria around innovation arising from our own experiences of WBAs. However, as a scoping review, we do consider we have identified sufficient studies to support the seven dimensions of the WBA tools we identified.

It is interesting to consider why there are so many tools available. Do these represent dispersed efforts to resolve common problems in assessing workplace performance, or are the perceived problems also localised? If the former, then we would seem to be witnessing a natural experiment, in which strategies may either eventually converge on similar solutions, or different successful strategies will emerge. Designers may manipulate the dimensions so that a WBA tool most effectively fulfils its role in an assessment programme.

It is also interesting to consider whether the different configurations of WBA dimensions discussed here have any practical impact on future performance of specialists. A gap identified in this review is the predictive value of the different assessment tools on future performance, a gap not unique to this review. Without the gold standard of comparable performance measures after graduation, we are left to fall back on principles and process measures. A potential for future research, to elucidate this gold standard, could be to collect data from the various WBA tools from practising specialists.

Innovations in WBA continue to emerge, with advances in information technology to support them. These innovations include new approaches to enhancing timely, action-oriented feedback[44] and use of patient data to generate assessment information[45] and entrustment-based discussion,[46] an innovation in CbD to bridge the gap between observed performance and predictions of future performance. We think the framework we have described would help readers to decide how they might tailor reported innovations to their unique contexts.

Conclusions

In synthesising the literature on WBA tool innovations, we have added a framework for categorising characteristics across seven dimensions. This framework systematically teases apart the ways in which tools can vary and the domains in which innovations are occurring. It also draws attention to the trade-offs inherent in tool design and selection, and

enables a more deliberate consideration of the tool characteristics most appropriate to the local context.

Authors' contributions

Study conception: JMW
Database search: TC
Selection/analysis of included articles: all authors
Collating of included articles: TC
Drafting of methods and results: TC
Drafting of paper: all authors
Critical review of paper: all authors
Writing of final version of paper: JMW
Approval of final version of paper: all authors
All authors agree to be accountable for all aspects of the work.

Declarations of interest

The authors declare that they have no conflicts of interest.

Funding

Australian and New Zealand College of Anaesthetists Research Foundation (S20/002).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bja.2021.06.038.

References

1. Harris P, Bhanji F, Topps M, et al. Evolving concepts of assessment in a competency-based world. *Med Teach* 2017; **39**: 603–8
2. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med Res Methodol* 2018; **18**: 1433
3. Cooke A, Smith D, Booth A. Beyond PICO: the SPIDER tool for qualitative evidence synthesis. *Qual Health Res* 2012; **22**: 1435–43
4. Ehrenfeld JM, McEvoy MD, Furman WR, Snyder D, Sandberg WS. Automated near-real-time clinical performance feedback for anesthesiology residents: one piece of the milestones puzzle. *Anesthesiology* 2014; **120**: 172–84
5. Xu VYY, Hamid J, von Maltzahn M, et al. Use of the consultation letter rating scale among geriatric medicine postgraduate trainees. *J Am Geriatr Soc* 2019; **67**: 2157–60
6. Acai A, Li S-A, Sherbino J, et al. Attending emergency physicians' perceptions of a programmatic workplace-based assessment system: the McMaster Modular Assessment Program (McMAP). *Teach Learn Med* 2019; **31**: 434–44
7. Chan TM, Sherbino J, McMAP Collaborators. The McMaster Modular Assessment Program (McMAP): a theoretically grounded work-based assessment system for an emergency medicine residency program. *Acad Med* 2015; **90**: 900–5
8. Chan TM, Sherbino J, Mercuri M. Nuance and noise: lessons learned from longitudinal aggregated assessment data. *J Grad Med Educ* 2017; **9**: 724–9
9. Cheung WJ, Dudek N, Wood TJ, Frank JR. Daily encounter cards—evaluating the quality of documented assessments. *J Grad Med Educ* 2016; **8**: 601–4
10. Regan L, Cope L, Omron R, Bright L, Bayram J. Do end-of-rotation and end-of-shift assessments inform Clinical Competency Committees' (CCC) decisions? *West J Emerg Med* 2018; **19**: 121–7
11. Braund H, Dalgarno N, McEwen L, Egan R, Reid MA, Baxter S. Involving ophthalmology departmental stakeholders in developing workplace-based assessment tools. *Can J Ophthalmol* 2019; **54**: 590–600
12. Emke AR, Park YS, Srinivasan S, Tekian A. Workplace-based assessments using pediatric critical care entrustable professional activities. *J Grad Med Educ* 2019; **11**: 430–8
13. Toprak A, Luhanga U, Jones S, Winthrop A, McEwen L. Validation of a novel intraoperative assessment tool: the surgical procedure feedback rubric. *Am J Surg* 2016; **211**: 369–76
14. Hicks PJ, Margolis MJ, Carraccio CL, et al. A novel workplace-based assessment for competency-based decisions and learner feedback. *Med Teach* 2018; **40**: 1143–50
15. Anderson JT. Assessments and skills improvement for endoscopists. *Best Pract Res Clin Gastroenterol* 2016; **30**: 453–71
16. Donato AA, Park YS, George DL, George DL, Schwartz A, Yudkowsky R. Validity and feasibility of the Minicard direct observation tool in 1 training program. *J Grad Med Educ* 2015; **7**: 225–9
17. Anderson CI, Basson MD, Ali M, et al. Comprehensive multicenter graduate surgical education initiative incorporating entrustable professional activities, continuous quality improvement cycles, and a web-based platform to enhance teaching and learning. *J Am Coll Surg* 2018; **227**: 64–76
18. Cooney CM, Redett 3rd RJ, Dorafshar AH, Zarrabi B, Lifchez SD. Integrating the NAS Milestones and handheld technology to improve residency training and assessment. *J Surg Educ* 2014; **71**: 39–42
19. Townend W, Gopal A, Flowerdew L, Farrow A, Crossley J. The Extended Supervised Learning Event (ESLE): assessing nontechnical skills in emergency medicine trainees in the workplace. *Ann Emerg Med* 2019; **74**: 670–8
20. Kumar S, Danino J, Skinner DW. The ward round assessment tool: a new workplace-based assessment tool. *Bull Royal Coll Surg Engl* 2013; **95**: 1–8
21. Danino J, Kumar S, Skinner D. The ward round assessment tool (WrAT)—a new work based assessment tool: F110. *Clin Otolaryngol* 2012; **37**: 17–72
22. Fitzpatrick R, Paterson NR, Watterson J, Seabrook C, Roberts M. Development and implementation of a mobile version of the O-SCORE assessment tool and case log for competency-based assessment in urology residency training: an initial assessment of utilization and acceptance among residents and faculty. *Can Urol Assoc J* 2019; **13**: 45–50
23. Smit MP, de Hoog M, Brackel HJL, Ten Cate O, Gemke RJBJ. A national process to enhance the validity of entrustment decisions for Dutch pediatric residents. *J Grad Med Educ* 2019; **11**: 158–64
24. Warm EJ, Held JD, Hellmann M, et al. Entrusting observable practice activities and milestones over the 36 months of an internal medicine residency. *Acad Med* 2016; **91**: 1398–405

25. van Bockel EAP, Walstock PA, van Mook WNKA, et al. Entrustable professional activities (EPAs) for postgraduate competency based intensive care medicine training in the Netherlands: the next step towards excellence in intensive care medicine training. *J Crit Care* 2019; **54**: 261–7
26. Yuan CM, Prince LK, Zwettler AJ, Nee R, Oliver 3rd JD, Abbott KC. Assessing achievement in nephrology training: using clinic chart audits to quantitatively screen competency. *Am J Kidney Dis* 2014; **64**: 737–43
27. Park YS, Zar FA, Norcini JJ, Tekian A. Competency evaluations in the next accreditation system: contributing to guidelines and implications. *Teach Learn Med* 2016; **28**: 135–45
28. Warrington S, Beeson M, Bradford A. Inter-rater agreement of end-of-shift evaluations based on a single encounter. *West J Emerg Med* 2017; **18**: 518–24
29. Rekman J, Hamstra SJ, Dudek N, Wood T, Seabrook C, Gofton W. A new instrument for assessing resident competence in surgical clinic: the Ottawa clinic assessment tool. *J Surg Educ* 2016; **73**: 575–82
30. Hanson JL, Rosenberg AA, Lane JL. Narrative descriptions should replace grades and numerical ratings for clinical performance in medical education in the United States. *Front Psychol* 2013; **4**: 668
31. Turner TL, Bhavaraju VL, Luciw-Dubas UA, et al. Validity evidence from ratings of pediatric interns and subinterns on a subset of pediatric milestones. *Acad Med* 2017; **92**: 809–19
32. Kameoka J, Kikukawa M, Kobayashi D, Okubo T, Ishii S, Kagaya Y. A medical record peer-review system to evaluate residents' clinical competence: criterion validity analysis. *Tohoku J Exp Med* 2019; **248**: 253–60
33. Van Heest AE, Agel J, Ames SE, et al. Resident surgical skills web-based evaluation: a comparison of 2 assessment tools. *J Bone Jt Surg Am* 2019; **101**: e18
34. Weller JM, Castanelli DJ, Chen Y, Jolly B. Making robust assessments of specialist trainees' workplace performance. *Br J Anaesth* 2017; **118**: 207–14
35. Weller JM, Misur M, Nicolson S, et al. Can I leave the theatre? A key to more reliable workplace-based assessment. *Br J Anaesth* 2014; **112**: 1083–91
36. O'Connor A, McCurtin A, Cantillon P, McGarr O. Illusions of specificity in power-laden clinical performance assessment. *Med Teach* 2018; **40**: 313–4
37. Van der Vleuten CPM, Schuwirth LWT, Driessen EW, et al. A model for programmatic assessment fit for purpose. *Med Teach* 2012; **34**: 205–14
38. Schut S, Driessen E, van Tartwijk J, van der Vleuten C, Heeneman S. Stakes in the eye of the beholder: an international study of learners' perceptions within programmatic assessment. *Med Educ* 2018; **52**: 654–63
39. Van der Vleuten CP, Schuwirth L, Driessen E, Govaerts MJB, Heeneman S. Twelve tips for programmatic assessment. *Med Teach* 2015; **37**: 641–6
40. Castanelli DJ, Weller JM, Molloy E, Bearman M. Shadow systems in assessment: how supervisors make progress decisions in practice. *Adv Health Sci Educ Theory Pract* 2020; **25**: 131–47
41. Castanelli DJ, Jowsey T, Chen Y, Weller JM. Perceptions of purpose, value, and process of the mini-Clinical Evaluation Exercise in anesthesia training. *Can J Anaesth* 2016; **63**: 1345–56
42. Gaunt A, Patel A, Rusius V, Royle TJ, Markham DH, Paulikowska T. 'Playing the game': how do surgical trainees seek feedback using workplace-based assessment? *Med Educ* 2017; **51**: 953–62
43. Konopasek L, Norcini J, Krupat E. Focusing on the formative: building an assessment system aimed at student growth and development. *Acad Med* 2016; **91**: 1492–7
44. Young JQ, McClure M. Fast, easy, and good. *Acad Med* 2020; **95**: 1546–9
45. Schumacher DJ, Martini A, Holmboe E, et al. Initial implementation of resident-sensitive quality measures in the pediatric emergency department: a wide range of performance. *Acad Med* 2020; **95**: 1248–55
46. ten Cate O, Schwartz A, Chen HC. Assessing trainees and making entrustment decisions: on the nature and use of entrustment-supervision scales. *Acad Med* 2020; **95**: 1662–9

Handling editor: Jonathan Hardman