Direct Observation Tools in Emergency Medicine: A Systematic Review of the Literature

Michael Gottlieb, MD[1], Jaime Jordan, MD[2], Jeffrey N. Siegelman, MD[3], Robert Cooney, MD, MSMedEd[4], Christine Stehman, MD[5], and Teresa M. Chan, MD, MHPE[6]

ABSTRACT

Objectives: Direct observation is important for assessing the competency of medical learners. Multiple tools have been described in other fields, although the degree of emergency medicine–specific literature is unclear. This review sought to summarize the current literature on direct observation tools in the emergency department (ED) setting.

Methods: We searched PubMed, Scopus, CINAHL, the Cochrane Central Register of Clinical Trials, the Cochrane Database of Systematic Reviews, ERIC, PsycINFO, and Google Scholar from 2012 to 2020 for publications on direct observation tools in the ED setting. Data were dual extracted into a predefined worksheet, and quality analysis was performed using the Medical Education Research Study Quality Instrument.

Results: We identified 38 publications, comprising 2,977 learners. Fifteen different tools were described. The most commonly assessed tools included the Milestones (nine studies), Observed Structured Clinical Exercises (seven studies), the McMaster Modular Assessment Program (six studies), Queen's Simulation Assessment Test (five studies), and the mini-Clinical Evaluation Exercise (four studies). Most of the studies were performed in a single institution, and there were limited validity or reliability assessments reported.

Conclusions: The number of publications on direct observation tools for the ED setting has markedly increased. However, there remains a need for stronger internal and external validity data.

Direct observation involves observation of the learner in the clinical or simulated setting, generating information which can then be utilized both to provide real-time formative feedback and to generate data for global assessments of the learner.[1] Direct observation is a commonly used method for assessment of medical trainees and is especially important in today's age of competency-based medical education (CBME).[2–4] It can provide essential information regarding a trainee's knowledge, behavior, and skills related to a particular context or environment. In addition to providing information for both formative and summative feedback, direct observation can also aid in deliberate practice, which is essential for developing expertise.[5] Direct observation of a trainee's skills is an essential component of a workplace-based assessment

From the [1]Department of Emergency Medicine, Rush University Medical Center, Chicago, IL; the [2]Department of Emergency Medicine, Ronald Reagan UCLA Medical Center, Los Angeles, CA; the [3]Department of Emergency Medicine, Emory University School of Medicine, Atlanta, GA; the [4]Department of Emergency Medicine, Geisinger Medical Center, Danville, PA; [5]South Bend Emergency Physicians, South Bend, IN; and the [6]Department of Medicine, Division of Emergency Medicine, McMaster University, Hamilton, Ontario, Canada.
Received June 3, 2020; revision received July 31, 2020; accepted August 9, 2020.
The authors have no relevant financial information or potential conflicts to disclose.
Author contributions: MG, JJ, JNS, RC, CS, and TMC all contributed to the study concept and design, acquisition of the data, analysis and interpretation of the data, drafting of the manuscript, and critical revision of the manuscript for important intellectual content.
Supervising Editor: Sally Santen, MD, PhD.
Address for correspondence and reprints: Michael Gottlieb, MD; e-mail: michaelgottliebmd@gmail.com.
AEM EDUCATION AND TRAINING 2021;5:1–17

doi: 10.1002/aet2.10519

program and thereby plays a key role in both education and advancement decisions.[6,7] This became particularly important since the Accreditation Council for Graduate Medical Education (ACGME) created the Next Accreditation System (also known as the Milestones) in 2012.[8]

At that time, while there were a variety of tools available for direct observation of clinical skills, very few had been evaluated in the emergency department (ED) environment.[7] The unique practice environment of the ED poses additional challenges for emergency medicine (EM) educators and program leadership.[9,10] Several of these challenges, including the feasibility of conducting direct observations amid patient care and supervision of acutely ill individuals, were discussed as part of a breakout session on assessment of observable learner performance in EM during the 2012 *Academic Emergency Medicine* (*AEM*) Consensus Conference on Education Research.[1] The resulting article from this breakout session identified several strategies for assessing learner performance, including both direct observation strategies and indirect approaches (e.g., resident portfolios, procedure logs, self-reflection).[1] The authors then highlighted the strengths, weakness, relative costs, and available outcome data of direct observations and suggested a research agenda to address gaps in knowledge (Figure 1).[1] The field of education research within EM has advanced substantially since that time.[11] However, it is unclear to what degree this call to action has been answered by EM. Therefore, it is important to understand the current evidence to inform future research efforts and best practices. The objective of this article is to perform a systematic review of the literature on direct observation tools in EM published since 2012.

METHODS

Our study follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for systematic reviews and was performed in accordance with best practice guidelines (Data Supplement S1, Appendix S1, available as supporting information in the online version of this paper, which is available at http://onlinelibrary.wiley.com/doi/10.1002/aet2.10519/full).[12] In conjunction with a medical librarian, we conducted a search of PubMed, Scopus, the Cumulative Index to Nursing and Allied Health Literature (CINAHL), the Cochrane Central Register of Clinical Trials, the Cochrane Database of Systematic Reviews, Education Resources Information Center (ERIC), PsycINFO, and Google Scholar to include citations from January 1, 2012, to January 27, 2020. Details of the search strategy are included in Data Supplement S1, Appendix S2. After completing our initial search, we then performed a targeted search of each identified direct observation tool combined with "emergency medicine" in PubMed to identify any potentially missed articles. We specifically focused on articles published since 2012 because this was the date of the *AEM* Consensus Conference, which included a focus on direct observation tools.[1] As such, we sought to identify new literature on direct observational tools within EM since that time period. We also reviewed the bibliographies of all included studies and review articles for potentially missed studies. Finally, we consulted with topic experts to help identify any further relevant studies. No funding was provided for this review.

Inclusion and Exclusion Criteria

This review sought to summarize the existing direct observational tools used for the evaluation of medical students and residents in the ED setting. Inclusion criteria included all articles directly describing direct observational tools, which could include descriptive studies, retrospective studies, prospective studies, and randomized controlled trials. Studies could be performed in the clinical ED setting or a simulated environment. We excluded narrative reviews, studies focused on other specialties, studies focusing exclusively on procedural skills, or studies where the authors did not address direct observation tools. We intentionally excluded studies focusing exclusively on procedural skills because these tools have a different focus than clinical skills assessment tools. We also prospectively planned to exclude studies not published in English or Spanish if there was no translated version available, although none were identified in our literature search.

Two investigators (MG, JJ) independently assessed studies for eligibility based upon the above criteria. All abstracts meeting the initial criteria were reviewed as full-text articles. Studies deemed to meet the eligibility criteria on full-text review were included in the final data analysis. Any discrepancies were resolved by in depth discussion and negotiated consensus.

Data Collection and Processing

Two investigators (JNS, RC) underwent initial training and extracted data into a predesigned data collection

1. Determine the number of direct observation assessments and types of patient encounters (e.g., critical diagnoses, chief complaints, diagnostic complexity) that are needed to provide a valid reflection of patient care competence for an individual resident.

2. Design and codify a process to create reliable and valid simulation, objective structured clinical, and oral examination assessments that use checklists (time to event or critical action) and global ratings to assess competence in ways that reflect expert clinical practice (which may use shortcuts) rather than simply the accomplishment of basic task lists.

3. Determine the number of global assessments needed to compose a valid assessment of a resident's patient care competence accounting for the known biases of this method.

4. Assess the validity and relevance of non-clinician evaluations in patient care competence given the influence of potential confounders.

5. Determine the validity of clinical metrics relative to other more-studied forms of assessment with good reliability and validity such as direct observation, OSCE, and simulation.

6. Develop standardized training programs and assessments for procedural skill acquisition (such as those for central line insertion), starting with no-risk methods such as simulated, cadaveric, or OSCE experiences and concluding with direct observation assessment during actual patient care and correlation to complications and patient outcomes.

Figure 1. 2012 *AEM* Consensus Conference on Education Research Agenda on Clinical Skills Assessment Tools[1].

form. The following information was abstracted: first author name, year of publication, study title, number of participants, study country, study location (e.g., clinical, simulation laboratory), study design (e.g., qualitative, retrospective, prospective, randomized controlled trial), learner population (e.g., medical student, resident, year of training), medical specialty of the learners, tool utilized, assessor training, assessor calibration, outcome(s) measured, and the main study findings. Given the significant clinical heterogeneity of the studies, a meta-analysis was not planned.

Quality Analysis

Two investigators (MG, JJ) underwent initial training and independently performed quality analysis using the Medical Education Research Study Quality Instrument (MERSQI).[13] The MERSQI is a 10-item tool (18 maximum points), which was specifically designed for evaluating medical education research and has been demonstrated to have good inter-rater reliability.[13] The investigators compared responses and resolved any discrepancies by in depth discussion and negotiated consensus. Traditionally, the MERSQI tool has a maximum score of 18 points. However, because some of the components of this tool will only apply to quantitative studies, it may artificially underscore qualitative studies. To address this, we adjusted the maximum possible points for qualitative studies to 10 (Data Supplement S1, Appendix S3).

RESULTS

Summary of Findings

The search identified 728 articles. After duplicates were removed, 634 studies were screened using titles and abstracts with 71 selected for full-text review (Figure 2). Of these, 38 studies (n = 2,977 learners) met inclusion criteria and are discussed further below (Table 1).

Twenty studies were conducted in the United States,[14–33] 13 were performed in Canada,[34–46] three took place in Australia,[47–49] and two were conducted in Taiwan.[50,51] Thirty-three studies involved resident

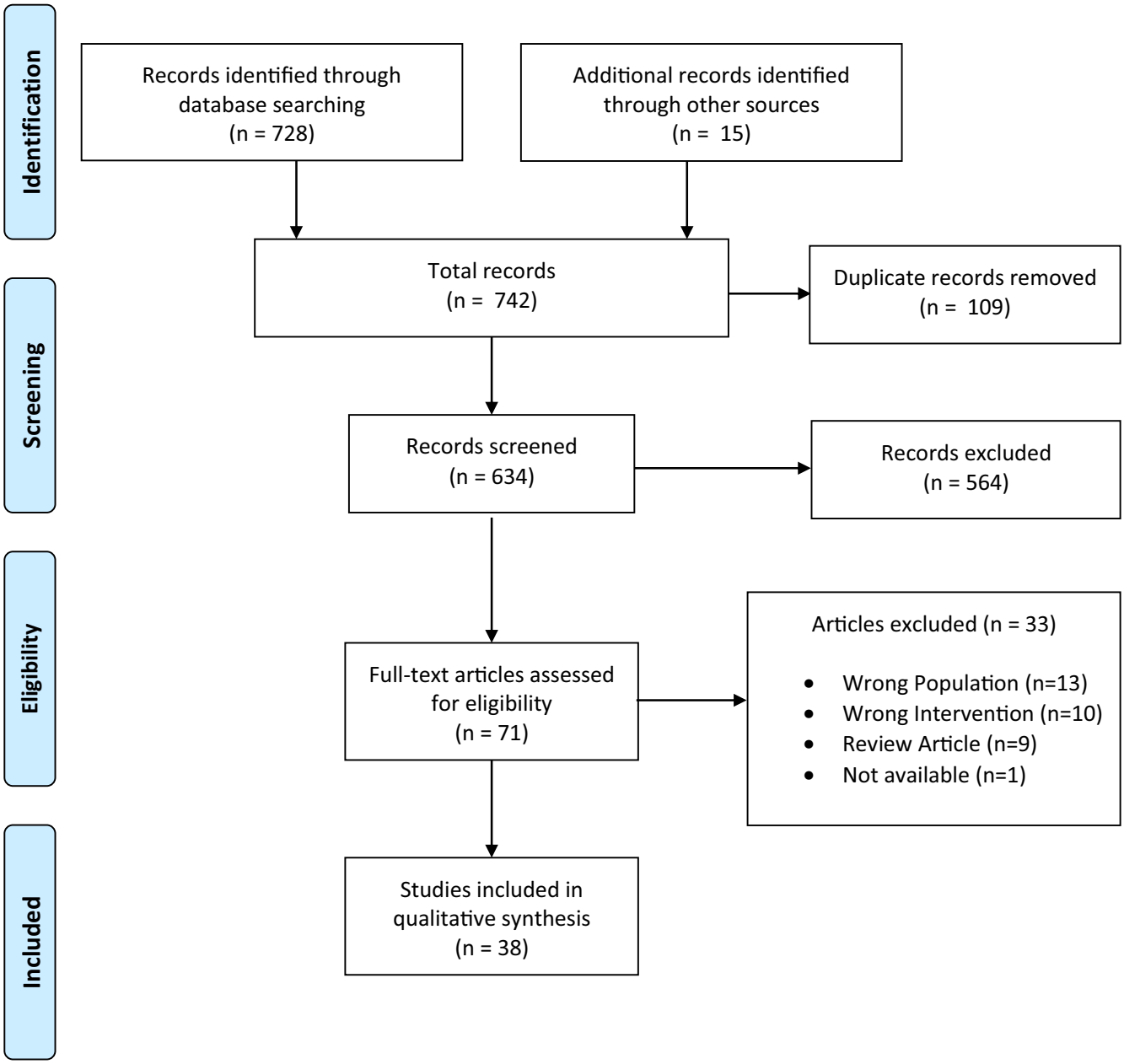

Figure 2. Flow diagram for article selection.

physicians,[15,17–23,25–32,34–47,49,50] four studies involved medical students,[14,16,24,33] and one did not report the learner level.[48] Twenty-two were prospective studies,[14,16,17,22–28,30,31,33,37,38,40–42,46–48,51] nine were retrospective,[18–20,32,35,39,44,45,50] five were qualitative studies,[21,29,34,43,49] and two were descriptive studies.[15,36] Twenty-five studies were performed in the clinical environment,[14,15,18–21,24,26–31,34–37,43–45,47–51] 12 in the simulation lab,[16,17,22,25,32,33,38–42,46] and one used both simulation and the clinical environment.[23]

Assessor training was described in 15 studies and ranged from a 10-minute training video to a dedicated 3-hour training session (Data Supplement S1, Appendix S4). Assessor calibration was described in nine studies and primarily consisted of either an initial calibration session or feedback based on subsequent scoring (Data Supplement S1, Appendix S4). Among studies assessing diagnostic accuracy, the most common criterion standard was training level (n = 9),[18–20,22,30,37–40] followed by independent scoring from experts (n = 3),[25,26,42] clinical competency committee scores (n = 2),[15,19] overall ED rotation score (n = 2),[14,33] in-training assessment report (n = 2),[41,47] and mean entrustment score from an alternate workplace-based assessment tool (n = 1).[46]

Table 1
Study Characteristics

Study	Number of Participants	Country	Study Location	Study Design	Learner Population	Tool Used
Acai 2019[34]	16	Canada	Clinical	Qualitative	PGY 1–5	McMAP
Ander 2012[14]	289	USA	Clinical	Prospective	MS 4	RIME and Global Rating Scale
Bedy 2019[15]	120	USA	Clinical	Descriptive	PGY 1–3	Milestones
Bord 2015[16]	80	USA	Simulation	Prospective	MS 2–4	OSCE, Milestones
Brazil 2012[47]	20	Australia	Clinical	Prospective	PGY 1	Mini-CEX
Bullard 2018[17]	30	USA	Simulation	Prospective	PGY 1	OSCE
Chan 2015[36]	15	Canada	Clinical	Descriptive	PGY 1, 2	McMAP
Chan 2017[35]	23	Canada	Clinical	Retrospective	PGY 2	McMAP
Chang 2017[50]	273	Taiwan	Clinical	Retrospective	PGY 1	Mini-CEX
Cheung 2019[37]	45	Canada	Clinical	Prospective	PGY 1–5	O-EDShOT
Dagnone 2016[38]	98	Canada	Simulation	Prospective	PGY 1–5	QSAT
Dayal 2017[18]	359	USA	Clinical	Retrospective	PGY 1–3	Milestones
Dehon 2015[19]	33	USA	Clinical	Retrospective	PGY 1–4	Milestones
Donato 2015[20]	73	USA	Clinical	Retrospective	PGY 1–3	Minicard
Edgerley 2018[39]	57	Canada	Simulation	Retrospective	PGY 1–5	QSAT
FitzGerald 2012[21]	34	USA	Clinical	Qualitative	PGY 1	Checklists
Hall 2015[40]	92	Canada	Simulation	Prospective	PGY 1–5	QSAT
Hall 2017[41]	79	Canada	Simulation	Prospective	PGY 1–5	QSAT
Hart 2018[22]	118	USA	Simulation	Prospective	PGY 1–3	Checklist, Global Rating Scale, Milestones
Hauff 2014[23]	28	USA	Clinical and Simulation	Prospective	PGY 1	OSCE, Checklist, Global Rating Scale, Milestones
Hoonpongsimanont 2018[24]	45	USA	Clinical	Prospective	MS 4	Local EOS evaluation
Hurley 2015[42]	57	Canada	Simulation	Prospective	PGY 3–5 and attending physicians	OSCE
Jones 2016[48]	24	Australia	Clinical	Prospective	PGY 1–4	Local EOS evaluation
Jong 2018[25]	34	USA	Simulation	Prospective	PGY 2–4	QSAT
Kane 2017[26]	26	USA	Clinical	Prospective	PGY 3–4	SDOT
Lee 2019[49]	73	Australia, NZ	Clinical	Qualitative	ND	Mini-CEX
Lefebvre 2018[27]	41	USA	Clinical	Prospective	PGY 1–3	Milestones
Li 2017[43]	26	Canada	Clinical	Qualitative	PGY 1–5	McMAP
Lin 2012[51]	230	Taiwan	Clinical	Prospective	PGY 1	Mini-CEX
McConnell 2016[44]	9	Canada	Clinical	Retrospective	PGY 2	McMAP
Min 2016[28]	10	USA	Clinical	Prospective	PGY 1–5	Global Breaking Bad News Assessment Scale
Mueller 2017[29]	71	USA	Clinical	Qualitative	PGY 3–4	Milestones
Paul 2018[30]	39	USA	Clinical	Prospective	PGY 1	OSCE, Checklist
Schott 2015[31]	29	USA	Clinical	Prospective	PGY 1–4	CDOT, Milestones
Sebok-Syer 2017[45]	23	Canada	Clinical	Retrospective	PGY 1–2	McMAP
Siegelman 2018[32]	102	USA	Simulation	Retrospective	PGY 1–3	OSCE
Wallenstein 2015[33]	239	USA	Simulation	Prospective	MS 4	OSCE
Weersink 2019[46]	17	Canada	Simulation	Prospective	PGY 1–5	RAT

CDOT = Critical Care Direct Observation Tool; EOS = end-of-shift; ITER = McMAP = McMaster Modular Assessment Program; Mini-CEX = Mini-Clinical Evaluation Exercise for Trainees; MS = medical student; ND = not described; NZ = New Zealand; O-EDShOT = Ottawa ED Shift Observation Tool; PGY = Post-graduate year; RAT = Resuscitation Assessment Tool; RIME = Reporter, Interpreter, Manager, Educator; SDOT = Standardized Direct Observation Tool.

Study Quality

The mean MERSQI score among the included studies was 13.1 of 18 (range = 8–15.5) for quantitative studies and 8.2 of 10 (range = 7–10) for qualitative studies (Table 2). The most common areas that studies lost points were study design, institutional

Table 2
Quality Assessment

Study	Study Design	Sampling: Institutions	Sampling: Response Rate	Type of Data	Validity Evidence: Content	Validity Evidence: Internal Structure	Validity Evidence: Relationship to Other Variables	Data Analysis: Sophistication	Data Analysis: Appropriate	Outcome	Total
Acai 2019[34]*	1	1.5	1	1	0	0	0	1	1	1	7.5
Ander 2012[14]	1	0.5	0.5	3	1	0	1	2	1	2	12
Bedy 2019[15]	1	0.5	1.5	3	1	0	1	1	1	2	12
Bord 2015[16]	1	0.5	1.5	3	1	1	1	2	1	1.5	13.5
Brazil 2012[47]	1	0.5	1.5	3	1	1	1	2	1	2	14
Bullard 2018[17]	1.5	0.5	1.5	3	1	1	1	2	1	1.5	14
Chan 2015[36]	1.5	0.5	0.5	3	1	0	1	2	1	2	12.5
Chan 2017[35]	1.5	0.5	1.5	3	1	0	1	2	1	2	13.5
Chang 2017[50]	1	0.5	0.5	3	1	1	1	2	1	2	13
Cheung 2019[37]	1	0.5	1.5	3	1	1	1	2	1	2	14
Dagnone 2016[38]	1	1.5	1.5	3	1	1	1	2	1	1.5	14.5
Dayal 2017[18]	1	1.5	1.5	3	1	0	1	2	1	2	14
Dehon 2015[19]	1	0.5	1.5	3	1	0	1	2	1	2	13
Donato 2015[20]	1	0.5	1.5	3	1	1	1	2	1	2	14
Edgerley 2018[39]	2	0.5	1.5	3	1	1	1	2	1	1.5	14.5
FitzGerald 2012[21]*	1	0.5	1.5	3	1	0	0	1	1	1	10
Hall 2015	1	0.5	1.5	3	1	1	1	2	1	1.5	13.5
Hall 2017	1	1.5	1.5	3	1	1	1	2	1	1.5	14.5
Hart 2018[22]	1	1.5	1.5	3	1	1	1	2	1	1.5	14.5
Hauff 2014[23]	1	0.5	1.5	3	1	1	1	1	1	2	13
Hoonpongsimanont 2018[24]	1.5	0.5	1.5	1	1	0	1	2	1	2	11.5
Hurley 2015[42]	1	0.5	1.5	3	1	1	1	2	1	1.5	13.5
Jones 2016[48]	1.5	0.5	1	1	1	0	0	1	1	1	8
Jong 2018[25]	1	0.5	1.5	3	1	1	1	2	1	1.5	13.5
Kane 2017[26]	1.5	0.5	1	3	1	0	0	2	1	1.5	11.5
Lee 2019[49]*	1	0.5	0.5	1	1	0	0	1	1	1	7
Lefebvre 2018[27]	1	0.5	1.5	3	1	0	0	2	1	2	12
Li 2017[43]*	1	0.5	1	1	1	0	0	1	1	1	7.5
Lin 2012[51]	1	1.5	1.5	3	1	0	0	1	1	2	12
McConnell 2016[44]	1	0.5	1.5	3	1	0	1	2	1	2	13

(Continued)

Table 2 (continued)

Study	Study Design	Sampling: Institutions	Sampling: Response Rate	Type of Data	Validity Evidence: Content	Validity Evidence: Internal Structure	Validity Evidence: Relationship to Other Variables	Data Analysis: Sophistication	Data Analysis: Appropriate	Outcome	Total
Min 2016[28]	1	0.5	1.5	3	1	0	1	2	1	2	13
Mueller 2017[29]*	1	0.5	1.5	1	1	0	0	1	1	2	9
Paul 2018[30]	2	1	1.5	3	1	1	1	2	1	2	15.5
Schott 2015[31]	1	1.5	1.5	3	1	1	1	2	1	1.5	14.5
Sebok-Syer 2017[45]	1	0.5	1.5	3	1	0	1	2	1	2	13
Siegelman 2018[32]	1	0.5	1.5	3	1	0	0	2	1	1.5	11.5
Wallenstein 2015[33]	1	0.5	1.5	3	1	0	1	2	1	1.5	12.5
Weersink 2019[46]	1	0.5	1	3	1	1	1	2	1	2	13.5

*Qualitative studies where the modified MERSQI tool was used.

sampling, and validity evidence for internal structure of the tools.

Direct Observation Tools

Studies evaluated 15 different direct observation tools, with the majority of the literature focusing on the following five tools. Ten studies utilized the ACGME EM Milestones,[15,16,18,19,22,23,27–29,31] seven used Observed Structured Clinical Exercises (OSCE),[16,17,23,30,32,33,42] six studies utilized the McMaster Modular Assessment Program (McMAP),[34–36,43–45] five used the Queen's Simulation Assessment Test (QSAT),[25,38–41] and four utilized the mini-Clinical Evaluation Exercise (mini-CEX).[47,49–51] Additional tools included checklists,[22,23,30] a global rating scale,[14,22,23] the Minicard,[20] a non–milestone-based end-of-shift evaluation,[24,47] the Ottawa Emergency Department Shift Observation Tool,[37] the Reporter/Interpreter/Manager/Educator (RIME) framework,[14] the Standardized Direct Observation Tool (SDOT),[26] the Critical Care Direct Observation Tool (CDOT),[31] and the Resuscitation Assessment Tool (RAT).[46] A summary of the data for each tool is provided in Table 3.

Milestone-based Evaluations

The tool most heavily represented in our sample was the ACGME EM Milestones. The Milestones are a framework for assessing resident progress, which were developed by each specialty to address the six core competencies created by the ACGME.[52] Multiple authors urged caution in using ACGME milestones to create end-of-shift or simulation assessment tools.[19,31] The CDOT, which allows for a milestone-based assessment of a resident during the early part of a critical resuscitation, did not demonstrate good reliability, with significant variability between raters (intraclass correlation from −0.04 to 0.019).[31] Dehon et al.[19] showed poor agreement at one site between end-of-shift milestone scores and clinical competency committee ratings, as well as similar rates of attainment of level 3 milestones for all resident levels. Alternatively, Dayal et al.[18] found that milestone scores increased 0.52 levels per year. Lefebvre et al.[27] found that including narrative comments along with milestone scores on end-of-shift tools increased the learner assessment scores assigned by the clinical competency committee.

Table 3
Summary of the Data for Each Tool

Direct Observation Tool	Total Studies (Total Participants)	AssessmentSetting	Accuracy and Reliability	Benefits	Limitations	Resource for Example Tool
CDOT	1 (29)	Clinical	Poor inter-rater reliability (ICC = −0.04 to 0.25)[31]	Focused on critical care interventions. Mapped to milestones. Includes a qualitative comments box.	Limited to yes, no, or N/A responses. Poor inter-rater reliability.	Schott 2015[31]
Checklists	4 (219)	Clinical and Simulation	Statistically significant increase for each training level (0.52 levels per year; $p < 0.001$).[22] Good inter-rater reliability (ICC = 0.81 to 0.86).[22]	Checklists are targeted to each clinical presentation. May include an area for qualitative feedback. If mapped to milestones, can also be used to evaluate milestones for ACGME.	Each checklist needs to be individually designed for each chief complaint. Primarily focused on specific presentations or aspects of care. Responseoptions often limited to yes, no, or unclear. Qualitative comments vary by checklist.	FitzGerald 2012[21] Hart 2018[22] Paul 2018[30]
Global Breaking Bad News Assessment Scale	1 (10)	Clinical	Resident skill increased by 90% on subsequent encounter.[28]	Short and easy to complete. Study tool can be modified to include a qualitative comments box.[28]	Only assesses delivery of bad news. Responses limited to yes or no.	Schildmann 2012[78]
Global Rating Scale	3 (435)	Clinical and Simulation	Statistically significant increase for each training level ($p < 0.05$).[22] Good inter-rater reliability for clinical management (ICC = 0.74 to 0.87) and communication (ICC = 0.80).[22]	Fewer questions. Faster to perform. Can be combined with other direct observation tools.	Relies heavily on gestalt. Less granular assessment of components. No qualitative comments.	Ander 2012[14] Hart 2018[22]
Local EOS Evaluation	2 (69)	Clinical	N/A	Can includes assessment of technical skills and some non-technical skills (e.g., professionalism, interpersonal skills)	Categorizations are general with limited specific examples. Not all tools have qualitative comments.	Hoonpongsimanont 2018[24] Jones 2016[48]
McMAP	6 (112)	Clinical	Data on accuracy not available. 12.7% variance between raters.[35]	Learner-centered. Individual clinical assessments were mapped to the ACGME and CanMEDS Frameworks. Tool uses behaviorally anchored scales and includes mandatory written comments.	May have a higher learning curve associated with the 76 unique assessments within the tool. Some components may not be possible to observe depending upon the patients encountered. Learners may avoid cumbersome tasks or those that they are weaker in. Faculty may avoid certain components that are harder to evaluate.	Acai 2019[34] Chan 2015[36] Chan 2017[35]

(Continued)

Table 3 (continued)

Direct Observation Tool	Total Studies (Total Participants)	AssessmentSetting	Accuracy and Reliability	Benefits	Limitations	Resource for Example Tool
Milestones	9 (911)	Clinical and Simulation	Statistically significant increase for each training level (0.52 levels per year; $p < 0.001$).[18] However, faculty may overestimate skills with milestones (92% milestone achievement regardless of training level).[19] Mean CCC score differed significantly from milestone scores ($p < 0.001$).[19] Poor inter-reliability in one study (ICC = −0.04 to 0.019).[31]	Addresses a diverse array of technical and nontechnical skills. Already utilized for summative residency assessments that are collected by the ACGME	Many of the milestones may not be applicable for a given patient or shift. Has a risk of grade inflation.[19] No qualitative comments.	ACGME Milestones[52]
Mini-CEX	4 (596)	Clinical	Did not identify any underperformers that were not already identified by the Australian Resident Medical Officer Assessment Form.[47]	Includes assessment of technical skills and some nontechnical skills (e.g., professionalism, efficiency). Overall high satisfaction among both learners and assessors.[47] Includes a dedicated area for qualitative feedback (strengths and weaknesses).	Does not assess teaching, teamwork, or documentation. Focused on single patient encounters so unable to account for managing multiple patients. Some components may be skipped unless they are required for completion.[50]	Brazil 2012[47]
Minicard	1 (73)	Clinical	Minicard scores increased by 0.021 points per month of training ($p < 0.001$).[20]	Includes comments for each individual assessment item. Includes an action plan at the end.	Inclusion of trainee level in descriptors for scoring may bias results.	Donato 2015[20]
O-EDShOT	1 (45)	Clinical	Statistically significant increase for each training level ($p < 0.001$).[37] 38% variance noted in ratings between raters.[37] 13 forms needed for 0.70 reliability.[37] 33 forms needed for 0.80 reliability.[37]	Designed specifically for the ED setting with feedback from faculty and residents. Includes an area for qualitative feedback (strengths and weaknesses). Can be used regardless of treatment area (i.e., high, medium, low acuity).	Only evaluated in a single study.	Cheung 2019[37]
OSCE	7 (575)	Clinical and simulation	OSCE was positively correlated with ED performance score ($p < 0.001$).[33] Comparing 20-item OSCE with 40-item OSCE revealed no difference in accuracy (85.6% vs. 84.5%).[42] Variation in ICC	Can assess a wide range of factors, including technical and nontechnical skills. Bullard modeled their tool after the ABEM oral board categories.[17]	OSCEs may vary between sites. OSCEs typically need to be individually designed for each presentation.	Bullard 2018[17] Paul 2018[30] Wallenstein 2015[33]

(Continued)

Table 3 (continued)

Direct Observation Tool	Total Studies (Total Participants)	AssessmentSetting	Accuracy and Reliability	Benefits	Limitations	Resource for Example Tool
			between studies (ICC = 0.43 to 0.92).[17,42]			
QSAT	5 (360)	Simulation	Statistically significant increase between PGY 1/2 and PGY 3–5 ($p < 0.001$).[38,40] Mean score increased by 10% for each training year ($p < 0.01$).[39] QSAT total score was moderately correlated with in-training evaluation report score ($r = 0.341$; $p < 0.01$).[41] Moderate inter-rater reliability (ICC = 0.56 to 0.89).[25,38,40]	Provides a framework that can be customized to each specific case.	Each QSAT would need to be individually designed for each presentation. Studies limited to the simulation environment.	Hall 2015[40] Hall 2017[41] Jong 2018[25]
RAT	1 (17)	Simulation	RAT was positively correlated with entrustment scores ($r = 0.630$; $p > 0.01$).[46] Moderate inter-rater reliability (ICC = 0.585 to 0.653).[46]	Builds upon QSAT with entrustable professional activities targeted towards resuscitation management. Designed using a modified Delphi study with experts.	Only assesses resuscitation management. Limited data from a single study.	Weersink 2019[46]
RIME	1 (289)	Clinical	Positive correlation between RIME category and clinical evaluation score ($r^2 = 0.40$, $p < 0.01$).[14] Very weak correlation between RIME category and clinical examination score.[14]	Easy to use. Can be combined with other tools.	Only one study evaluated RIME in the ED. Limited assessment of professional competencies (e.g., work ethic, teamwork, humanistic qualities).	Ander 2012[14]
SDOT	1 (26)	Clinical	Attending physicians were 54.4% accurate and resident physicians were 49.6% accurate when compared with the criterion standard scoring.[26]	Includes assessment of technical skills and some nontechnical skills (e.g., professionalism, interpersonal skills).	Several components may not be applicable to some patient encounters. Does not include an option for qualitative comments. Lower accuracy compared with other tools. May be more time consuming than other direct observation tools.	Kane 2017[26]

ACGME = Accreditation Council for Graduate Medical Education; CCC = clinical competency committee; CDOT = Critical Care Direct Observation Tool; EOS = end of shift; ICC = intraclass correlation; McMAP = McMaster Modular Assessment Program; Mini-CEX = Mini-Clinical Evaluation Exercise for Trainees; N/A = not available; O-EDShOT = Ottawa ED Shift Observation Tool; RAT = Resuscitation Assessment Tool; RIME = Reporter, Interpreter, Manager, Educator; SDOT = Standardized Direct Observation Tool.

OSCE

Several authors examined various tools used for OSCEs. OSCEs are highly structured tools used to assess competency, with an emphasis on objective assessment measures.[53] Bord et al.[16] developed an OSCE to evaluate attainment of Level 1 milestones in clerkship students, which was able to discriminate between high- and low-performing students. Wallenstein and Ander[33] compared OSCE scores with overall EM clerkship scores and found that they were positively correlated ($p < 0.001$). Hauff et al.[23] described an OSCE developed as part of postgraduate orientation which showed that many of their incoming interns had not attained Level 1 milestones. Bullard et al.[17] created an OSCE modeled after the American Board of Emergency Medicine oral board examination and found high inter-rater reliability (intraclass correlation = 0.92). The authors also noted a retained educational benefit for both the participants and the observers of the OSCE at 3 months.[17] Hurley et al.[42] evaluated the effect of the OSCE length on interobserver reliability and accuracy and found that no significant difference was present between the 20-item and 40-item checklists.

McMAP

The McMAP is a competency-based program of assessment that includes 76 micro clinical assessments systematically mapped to key clinical tasks within EM.[34,35] Each specific task includes a checklist to orient the rater, a global assessment using behavioral anchors, and mandatory written comments.[34] Each clinical task is linked to a global end-of-shift rating.[34] Both resident and faculty reflections on the implementation of this tool have been published.[34,43] Key benefits described by faculty were the inclusion of a wide range of clinical tasks, learner-driven emphasis, and the facilitation of more targeted specific and global feedback.[34] However, the faculty also noted that there was a learning curve associated with the 76 unique assessment instruments.[34] There was also concern among faculty that learners could "game the system" by selecting tasks that they were more facile with.[34] Authors used the data from the McMAP implementation to comment on systematic gaps in data collection (such as the Health Advocate and Professional CanMEDS roles)[44] and the effect of McMAP on end-of-year report quality for residents,[36] analyze narrative comments compared with checklist scores,[45] and describe longitudinal patterns arising from aggregate assessments.[35] The McMAP tool increased resident perception of formative feedback delivery and provided a conduit for residents to seek real-time feedback.[36,43]

QSAT

The QSAT is a modification of the OSCE that incorporates a global learner assessment score and was shown to discriminate well between learner levels.[38,40] It has also demonstrated good inter-rater reliability with one study reporting an intraclass correlation coefficient (ICC) of 0.89,[38] while a different study reported individual ICCs ranging from 0.56 to 0.87.[40] It has been used to provide complementary data to an in-training evaluation report used in Canada, assessing different aspects of competence.[41] Scores have been demonstrated to increase 10% with each additional year of training.[39] Interestingly, Edgerley et al.[39] found that working a night shift within one day of a QSAT assessment did not significantly impact a learner's score.

Mini-CEX

The Mini-CEX was originally utilized in internal medicine and has more recently been adapted to the ED setting. The Mini-CEX is a direct observation tool which emphasizes the following domains: medical interviewing, physical examination, humanistic qualities/professionalism, clinical judgment, counseling, and organization/efficiency.[54] Lin et al.[51] showed that most raters using this tool focused on clinical judgment, with decreased emphasis on humanistic components, while Lee et al.[49] explored factors influencing rater judgments using the tool. Brazil et al.[47] found that the Mini-CEX increased formative feedback overall despite not addressing all of the performance domains. Chang et al.[50] demonstrated that Mini-CEX compliance was improved on a computer format, particularly among raters with less than 10 years of seniority.

Other Tools

A variety of other tools were evaluated in a more limited number of studies. Ander et al.[14] found that the RIME framework correlated well with the overall clinical evaluation score in an EM clerkship ($r^2 = 0.40$, $p < 0.01$). Kane et al.[26] evaluated a training session for the SDOT among attending and resident physicians, noting that even after training, the attending physicians selected the correct rating in only 54.4% of cases, while senior residents selected the correct rating in 49.6% of cases. One group described a set of four

different checklists that were each targeted to evaluating specific history and physical examination skills based on chief complaints (e.g., asthma, fever in the neonate, pediatric fever, and gastroenteritis/dehydration).[21] Hart et al.[22] studied global rating scales for clinical management and communications, as well as a checklist mapped to the milestones, and reported a statistically significant increase with level of training ($p < 0.001$). They also noted an ICC of 0.74 to 0.87 for the global rating scale and an ICC of 0.81 to 0.86 for the checklist.[22] Paul and colleagues[30] described an otoscopy skills-focused tool. More recently, the Ottawa Emergency Department Shift Observation Tool was introduced as an entrustment-based tool to evaluate a resident's ability to manage the ED, with some validity evidence supporting its use, but further studies are needed.[37] Jones and Nanda[48] reported on a locally developed workplace-based assessment tool which raters found useful. Weersink et al.[46] modified the QSAT to develop the RAT, which demonstrated a positive correlation with resident entrustment scores ($r = 0.630$, $p < 0.01$) and good inter-rater reliability (ICC = 0.59 to 0.65). Hoonpongsimanont et al.[24] described a locally developed workplace-based assessment tool used on learner performances recorded with GoogleGlass and compared this with learner self-assessments of that same recording. Donato et al.[20] evaluated a Minicard that was able to identify struggling learners and also encouraged formative feedback with action plans.

Bias in Assessment

Three studies evaluated the effect of sex on evaluation results. Dayal et al.[18] found that milestone-based assessments used at the point of care or end of shift led to a 12.7% higher score for males compared to females regardless of assessor sex or assessor–assessee pairing. This corresponded to approximately 3 to 4 months of additional training in their study.[18] Mueller et al.[29] found that female residents received less consistent feedback than their male counterparts. Feedback was particularly inconsistent regarding issues of autonomy and assertiveness.[29] Siegelman et al.[32] evaluated bias in simulation-based assessments but did not find a similar association between rater or trainee sex and score.

Multisource Feedback

Three studies specifically evaluated the use of non-physician observers.[15,25,28] Bedy et al.[15] utilized ED-based pharmacists to specifically evaluate EM resident performance of the Pharmacotherapy Milestone (PC5). They found that pharmacist input was valuable for the determination of milestone levels during the clinical competency committee meetings.[15] Jong et al.[25] studied the QSAT among residents comparing physicians with nurses and emergency medical technicians (EMTs) as raters. In their study, nurses had moderate agreement with physicians (ICC = 0.65), while EMTs had excellent agreement with physicians (ICC = 0.812) for the QSAT scoring.[25] Social worker evaluation of resident performance during the delivery of bad news was studied by Min et al.[28] They found that this was acceptable to both residents and social workers, but that social workers tended to rate resident performance higher when compared with the resident's self-assessment.[28]

DISCUSSION

Since the 2012 *AEM* Consensus Conference, there has been a substantial increase in the number of publications related to direct observation tools in the ED setting. This is encouraging, as Kogan et al.[7] identified only six total EM-based studies in their prior systematic review. We were able to identify 38 new studies since 2012 alone. This adds significantly to the available literature on this topic.

The most common tool utilized was the ACGME EM Milestones despite the intent that milestones would guide assessment practices instead of acting as the assessment.[55] This is not surprising, because this is utilized by all EM residency programs as part of their assessment of residents and is required to be reported to the ACGME for reaccreditation. Therefore, it would seem reasonable to extend these summative assessments to direct observation tools. However, studies found relatively limited reliability of the measurements when used as direct observation tools.[19,31] These findings may reflect a problem with the tools used or with how clinicians are trained to use them.[56] None of the studies included in our review discussed how the assessors were trained with regard to this assessment tool. Multiple studies have suggested that potential assessors need to be sufficiently trained (including targeted training sessions, initial calibration, and feedback on assessment scoring with regard to one's peers)[57–59] and that they need to see the value in improving their ability to assess their learners.[60–63]

Some of the more successful tools for assessing EM learner skills include the McMAP, QSAT, OSCEs, O-

EDShOT, RAT, global rating scale, and checklists. While these tools demonstrated greater discriminatory ability, they were often limited to a small group of learners in a single program. While the Working Group on Assessment of Observable Learner Performance emphasized the need to refine previously well-established tools (e.g., mini-CEX, SDOT),[64] nearly one-third of studies described a novel tool. Unfortunately, there is limited validity evidence for most of these tools. When creating a new tool, it is important to demonstrate the validity of the measure, and future studies should seek to better establish the internal and external validity of these newer tools. This should include assessment of content, response process, internal structure, relationship to other variables, and consequences.[64,65] Moreover, studies should ensure that they follow and explicitly report adherence with recommended reporting guidelines.[66]

The overall quality of the data was lower with mean MERSQI scores of 13.1 of 18 for quantitative studies and 8.2 of 10 for qualitative studies. Many studies lost several points for insufficient validity evidence as described above. Additionally, the vast majority of studies were single-group, cross-sectional analyses. As the evidence advances, there will be a need for more cohort and randomized controlled trials comparing different direct observation methods. Finally, while most studies were performed at a single site, there is a need to assess these interventions across multiple institutions to better evaluate external validity.

Within our data, we found conflicting information regarding the effect of sex on direct observation tools. Sex biases and inequality have been demonstrated in EM among attending physicians, but the data among EM residents are more limited.[67–71] Dayal et al.[18] found that female residents had a lower overall rating on end-of-shift evaluations compared with male residents, while Mueller et al.[29] found that female residents received less consistent feedback than male residents. Interestingly, Siegelman et al.[32] did not find a difference in scoring between male and female residents in their simulation-based OSCE study. This may be due to the use a simulation environment, where there was a greater degree of control and external supervision, which may have led to a Hawthorne effect. Additionally, the OSCE is a binary tool (yes/no), which may be less prone to bias than the more subjective tools. Further studies should assess the role of bias in assessment and strategies to prevent this. Additionally, studies should also evaluate the impact of other nonmedical biases related to race, ethnicity, age, and primary language.

When compared with the 2012 *AEM* Consensus Conference on Education Research Direct Observation Tools Research Agenda (Figure 1),[1] there remains a need for better data on the reliability and discriminatory ability of the assessment tools in the ED environment and among different assessors (Items 2 and 4). Three studies described the role of nonphysicians (e.g., pharmacists, social workers) using direct observation tools,[15,25,28] but none compared these assessments with a criterion standard. Additionally, there remains a need for more data on the number of direct observations, types of patient encounters, and global assessments necessary to determine competency (Items 1 and 3). The ED is a unique environment, wherein attending and resident physicians are in close proximity to each other for an extended period of time. However, time constraints may limit direct observation and variations in intervals between consecutive shifts may limit the ability to develop longitudinal assessments of progress.[72–75] Future studies need to determine the ideal number and distribution of encounters necessary to reliably assess competence. We did not identify any studies comparing the validity of clinical metrics relative to other forms observation (Item 5), so there remains a need for further research into this area. We did not assess procedural skill acquisition (Item 6) in our review.

As the field moves increasingly toward CBME, there will be an increased need for validity and reliability evidence for direct observation tools in the ED setting.[76,77] Future research will need to build on these studies to better assess validity across institutions and among different providers as well as how best to integrate this into the clinical ED environment.

LIMITATIONS

There are several limitations with regard to this study. First, the studies had substantial heterogeneity with regard to tools and outcome measures, limiting the ability to perform meta-analysis for any of the outcomes. Among the included studies, many utilized the tool as part of the overall assessment, which may have led to incorporation bias. Also, the criterion standard varied between studies with many using training level, which may not reflect the actual degree of clinical expertise. Only a limited number of studies adequately described assessor training and calibration. Future studies should ensure that these are fully described in the methods.

Additionally, the search was limited to EM and did not include direct observation tools developed and/or evaluated in other specialties that may be applicable to the ED environment. Moreover, only three studies evaluated medical students, so it remains unclear how reliable most of these tools are in this learner group. The majority of studies were performed in North America. As such, it is unclear how this would apply in other locations. Our search excluded procedural assessment tools, so we were not able to comment on this in the article. Finally, while we searched eight databases with the assistance of a medical librarian, it is possible that we may have missed some relevant studies. However, we also performed bibliographic review of all included articles and reached out to topic experts, so we believe that the risk of this is low.

CONCLUSION

There is a burgeoning body of work within emergency medicine focusing on how we might optimize direct observation of our trainees. The majority of these articles assess the Milestones, McMAP, OSCE, QSAT, and Mini-CEX, although validity evidence is limited. Future studies are needed to better assess the validity, reliability, and number of evaluations necessary to assess competence.

The authors would like to thank Jennifer C. Westrick, MSLIS for her assistance with the literature search.

REFERENCES

1. Takayesu JK, Kulstad C, Wallenstein J, et al. Assessing patient care: summary of the breakout group on assessment of observable learner performance. Acad Emerg Med 2012;19:1379–89.
2. Carraccio C, Wolfsthal SD, Englander R, Ferentz K, Martin C. Shifting paradigms: from Flexner to competencies. Acad Med 2002;77:361–7.
3. Fromme HB, Karani R, Downing SM. Direct observation in medical education: a review of the literature and evidence for validity. Mt Sinai J Med 2009;76:365–71.
4. Andolsek KM, Simpson D. Direct observation reassessed. J Grad Med Educ 2017;9:531–32.
5. Ericsson KA. Deliberate practice and acquisition of expert performance: a general overview. Acad Emerg Med 2008;15:988–94.
6. Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: asystematic review. BMJ 2010;341:c5064.
7. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. JAMA 2009;302:1316–26.
8. Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system–rationale and benefits. N Engl J Med 2012;366:1051–6.
9. Reiter M.Emergency Medicine: The Good, the Bad, and the Ugly. 2011. Available at: https://www.medscape.com/viewarticle/750482#vp_2. Accessed Jul 26, 2020,
10. Suter RE. Emergency medicine in the United States: a systemic review. World J Emerg Med 2012;3:5–10.
11. Gottlieb M, Chan TM, Clarke SO, et al. Emergency medicine education research since the 2012 consensus conference: how far have we come and what's next? AEM Educ Train 2019;4:S57–66.
12. Moher D, Liberati A, Tetzlaff J, Altman DG. PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6:e1000097.
13. Cook DA, Reed DA. Appraising the quality of medical education research methods: the Medical Education Research Study Quality Instrument and the Newcastle-Ottawa Scale-Education. Acad Med 2015;90:1067–76.
14. Ander DS, Wallenstein J, Abramson JL, Click L, Shayne P. Reporter-Interpreter-Manager-Educator (RIME) descriptive ratings as an evaluation tool in an emergency medicine clerkship. J Emerg Med 2012;43:720–7.
15. Bedy SC, Goddard KB, Stilley JA, Sampson CS. Use of emergency department pharmacists in emergency medicine resident milestone assessment. West J Emerg Med 2019;20:357–62.
16. Bord S, Retezar R, McCann P, Jung J. Development of an objective structured clinical examination for assessment of clinical skills in an emergency medicine clerkship. West J Emerg Med 2015;16:866–70.
17. Bullard MJ, Weekes AJ, Cordle RJ, et al. A mixed-methods comparison of participant and observer learner roles in simulation education. AEM Educ Train 2018;3:20–32.
18. Dayal A, O'Connor DM, Qadri U, Arora VM. Comparison of male vs female resident milestone evaluations by faculty during emergency medicine residency training. JAMA Intern Med 2017;177:651–7.
19. Dehon E, Jones J, Puskarich M, Sandifer JP, Sikes K. Use of emergency medicine milestones as items on end-of-shift evaluations results in overestimates of residents' proficiency level. J Grad Med Educ 2015;7:192–6.
20. Donato AA, Park YS, 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.
21. FitzGerald M, Mallory M, Mittiga M, et al. Experience-based guidance for implementing a direct observation

checklist in a pediatric emergency department setting. J Grad Med Educ 2012;4:521–4.
22. Hart D, Bond W, Siegelman JN, et al. Simulation for assessment of milestones in emergency medicine residents. Acad Emerg Med 2018;25:205–20.
23. Hauff SR, Hopson LR, Losman E, et al. Programmatic assessment of level 1 milestones in incoming interns. Acad Emerg Med 2014;21:694–8.
24. Hoonpongsimanont W, Feldman M, Bove N, et al. Improving feedback by using first-person video during the emergency medicine clerkship. Adv Med Educ Pract 2018;9:559–65.
25. Jong M, Elliott N, Nguyen M, et al. Assessment of emergency medicine resident performance in an adult simulation using a multisource feedback approach. West J Emerg Med 2019;20:64–70.
26. Kane KE, Weaver KR, Barr GC Jr, et al. Standardized direct observation assessment tool: using a training video. J Emerg Med 2017;52:530–7.
27. Lefebvre C, Hiestand B, Glass C, et al. Examining the effects of narrative commentary on evaluators' summative assessments of resident performance. Eval Health Prof 2020;43:159–61.
28. Min AA, Spear-Ellinwood K, Berman M, Nisson P, Rhodes SM. Social worker assessment of bad news delivery by emergency medicine residents: a novel direct-observation milestone assessment. Intern Emerg Med 2016;11:843–52.
29. Mueller AS, Jenkins TM, Osborne M, Dayal A, O'Connor DM, Arora VM. Gender differences in attending physicians' feedback to residents: a qualitative analysis. J Grad Med Educ 2017;9:577–85.
30. Paul CR, Keeley MG, Rebella GS, Frohna JG. Teaching pediatric otoscopy skills to pediatric and emergency medicine residents: a cross-institutional study. Acad Pediatr 2018;18:692–7.
31. Schott M, Kedia R, Promes SB, et al. Direct observation assessment of milestones: problems with reliability. West J Emerg Med 2015;16:871–6.
32. Siegelman JN, Lall M, Lee L, Moran TP, Wallenstein J, Shah B. Gender bias in simulation-based assessments of emergency medicine residents. J Grad Med Educ 2018;10:411–5.
33. Wallenstein J, Ander D. Objective structured clinical examinations provide valid clinical skills assessment in emergency medicine education. West J Emerg Med 2015;16:121–6.
34. Acai A, Li SA, Sherbino J, Chan TM. Attending emergency physicians' perceptions of a programmatic workplace-based assessment system: the McMaster modular assessment program (McMAP). Teach Learn Med 2019;31:434–44.
35. Chan TM, Sherbino J, Mercuri M. Nuance and noise: lessons learned from longitudinal aggregated assessment data. J Grad Med Educ 2017;9:724–9.
36. Chan T, 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.
37. Cheung WJ, Wood TJ, Gofton W, Dewhirst S, Dudek N. The Ottawa Emergency Department Shift Observation Tool (O-EDShOT): a new tool for assessing resident competence in the emergency department. AEM Educ Train 2019 [Online ahead of print]. https://doi.org/10.1002/aet2.10419
38. Dagnone JD, Hall AK, Sebok-Syer S, et al. Competency-based simulation assessment of resuscitation skills in emergency medicine postgraduate trainees - a Canadian multi-centred study. Can Med Educ J 2016;7:e57–67.
39. Edgerley S, McKaigney C, Boyne D, Ginsberg D, Dagnone JD, Hall AK. Impact of night shifts on emergency medicine resident resuscitation performance. Resuscitation 2018;127:26–30.
40. Hall AK, Dagnone JD, Lacroix L, Pickett W, Klinger DA. Queen's simulation assessment tool: development and validation of an assessment tool for resuscitation objective structured clinical examination stations in emergency medicine. Simul Healthc 2015;10:98–105.
41. Hall AK, Damon Dagnone J, Moore S, et al. Comparison of simulation-based resuscitation performance assessments with in-training evaluation reports in emergency medicine residents: a Canadian multicenter study. AEM Educ Train 2017;1:293–300.
42. Hurley KF, Giffin NA, Stewart SA, Bullock GB. Probing the effect of OSCE checklist length on inter-observer reliability and observer accuracy. Med Educ Online 2015;20:29242.
43. Li SA, Sherbino J, Chan TM. McMaster Modular Assessment Program (McMAP) through the years: residents' experience with an evolving feedback culture over a 3-year period. AEM Educ Train 2017;1:5–14.
44. McConnell M, Sherbino J, Chan TM. Mind the gap: the prospects of missing data. J Grad Med Educ 2016;8:708–12.
45. Sebok-Syer SS, Klinger DA, Sherbino J, Chan TM. Mixed messages or miscommunication? Investigating the relationship between assessors' workplace-based assessment scores and written comments. Acad Med 2017;92:1774–9.
46. Weersink K, Hall AK, Rich J, Szulewski A, Dagnone JD. Simulation versus real-world performance: a direct comparison of emergency medicine resident resuscitation entrustment scoring. Adv Simul (Lond) 2019;4:9.
47. Brazil V, Ratcliffe L, Zhang J, Davin L. Mini-CEX as a workplace-based assessment tool for interns in an emergency department–does cost outweigh value? Med Teach 2012;34:1017–23.
48. Jones CL, Nanda R. Assessing a doctor you've rarely worked with: the use of workplace-based assessments in a

busy inner city emergency department. Emerg Med Australas 2016;28:439–43.
49. Lee V, Brain K, Martin J. From opening the 'black box' to looking behind the curtain: cognition and context in assessor-based judgements. Adv Health Sci Educ Theory Pract 2019;24:85–102.
50. Chang YC, Lee CH, Chen CK, et al. Exploring the influence of gender, seniority and specialty on paper and computer-based feedback provision during mini-CEX assessments in a busy emergency department. Adv Health Sci Educ Theory Pract 2017;22:57–67.
51. Lin CS, Chiu TF, Yen DH, Chong CF. Mini-clinical evaluation exercise and feedback on postgraduate trainees in the emergency department: a qualitative content analysis. J Acute Med 2012;2:1–7.
52. Accreditation Council for Graduate Medical Education. The Emergency Medicine Milestones Project. 2015. Available at: https://www.acgme.org/Portals/0/PDFs/Milestones/EmergencyMedicineMilestones.pdf?ver=2015-11-06-120531-877. Accessed Jul 26, 2020.
53. Harden RM. What is an OSCE? Med Teach 1988;10:19–22.
54. Mortaz Hejri S, Jalili M, Masoomi R, Shirazi M, Nedjat S, Norcini J. The utility of mini-Clinical Evaluation Exercise in undergraduate and postgraduate medical education: a BEME review: BEME Guide No. 59. Med Teach 2020;42:125–42.
55. Carter WA Jr. Milestone myths and misperceptions. J Grad Med Educ 2014;6:18–20.
56. Sheng AY. Trials and tribulations in implementation of the emergency medicine milestones from the frontlines. West J Emerg Med 2019;20:647–50.
57. Stefan A, Hall JN, Sherbino J, Chan TM. Faculty development in the age of competency-based medical education: a needs assessment of Canadian emergency medicine faculty and senior trainees. CJEM 2019;21:527–34.
58. Cheung WJ, Patey AM, Frank JR, Mackay M, Boet S. Barriers and enablers to direct observation of trainees' clinical performance: a qualitative study using the theoretical domains framework. Acad Med 2019;94:101–14.
59. Kilbertus S, Pardhan K, Zaheer J, Bandiera G. Transition to practice: evaluating the need for formal training in supervision and assessment among senior emergency medicine residents and new to practice emergency physicians. CJEM 2019;21:418–26.
60. Hodwitz K, Kuper A, Brydges R. Realizing one's own subjectivity: assessors' perceptions of the influence of training on their conduct of workplace-based assessments. Acad Med 2019;94:1970–9.
61. Favreau MA, Tewksbury L, Lupi C, et al. Constructing a shared mental model for faculty development for the core entrustable professional activities for entering residency. Acad Med 2017;92:759–64.
62. Kogan JR, Conforti LN, Yamazaki K, Iobst W, Holmboe ES. Commitment to change and challenges to implementing changes after workplace-based assessment rater training. Acad Med 2017;92:394–402.
63. Kogan JR, Conforti LN, Bernabeo E, Iobst W, Holmboe E. How faculty members experience workplace-based assessment rater training: a qualitative study. Med Educ 2015;49:692–708.
64. Kessler CS, Leone KA. The current state of core competency assessment in emergency medicine and a future research agenda: recommendations of the working group on assessment of observable learner performance. Acad Emerg Med 2012;19:1354–9.
65. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ 2003;37:830–7.
66. EQUATOR Network.Reporting Guidelines. Available at: https://www.equator-network.org/reporting-guidelines/. Accessed Jul 26, 2020.
67. Bennett CL, Raja AS, Kapoor N, et al. Gender differences in faculty rank among academic emergency physicians in the United States. Acad Emerg Med 2019;26:281–5.
68. Krzyzaniak SM, Gottlieb M, Parsons M, Rocca N, Chan TM. What emergency medicine rewards: is there implicit gender bias in national awards? Ann Emerg Med 2019;74:753–8.
69. Gottlieb M, Krzyzaniak SM, Mannix A, et al. Sex distribution of editorial board members among emergency medicine journals. Ann Emerg Med 2020 [Online ahead of print]. https://doi.org/10.1016/j.annemergmed.2020.03.027
70. Mannix A, Parsons M, Krzyzaniak SM, et al. Emergency Medicine Gender in Resident Leadership Study (EM GIRLS): the gender distribution among chief residents. AEM Educ Train 2020;4:262–5.
71. Brucker K, Whitaker N, Morgan ZS, et al. Exploring gender bias in nursing evaluations of emergency medicine residents. Acad Emerg Med 2019;26:1266–72.
72. Burdick WP, Schoffstall J. Observation of emergency medicine residents at the bedside: how often does it happen? Acad Emerg Med 1995;2:909–13.
73. Chisholm CD, Whenmouth LF, Daly EA, Cordell WH, Giles BK, Brizendine EJ. An evaluation of emergency medicine resident interaction time with faculty in different teaching venues. Acad Emerg Med 2004;11:149–55.
74. Flowerdew L, Brown R, Vincent C, Woloshynowych M. Development and validation of a tool to assess emergency physicians' nontechnical skills. Ann Emerg Med 2012;59:376–85.
75. Buckley C, Natesan S, Breslin A, Gottlieb M. Finessing feedback: recommendations for effective feedback in the emergency department. Ann Emerg Med 2020;75:445–51.

76. McGaghie WC, Miller GE, Sajid AW, Telder TV. Competency-based curriculum development on medical education: an introduction. Public Health Pap 1978; 68:11–91.
77. Van Melle E, Frank JR, Holmboe ES, et al. A core components framework for evaluating implementation of competency-based medical education programs. Acad Med 2019;94:1002–9.
78. Schildmann J, Kupfer S, Burchardi N, Vollmann J. Teaching and evaluating breaking bad news: a pre-post evaluation study of a teaching intervention for medical students and a comparative analysis of different measurement instruments and raters. Patient Educ Couns 2012;86: 210–19.

Supporting Information

The following supporting information is available in the online version of this paper available at http://onlinelibrary.wiley.com/doi/10.1002/aet2.10519/full

Data Supplement S1. Supplementary materials.