Tools for Direct Observation and Assessment of Clinical Skills of Medical Trainees: A Systematic Review

Article *in* JAMA The Journal of the American Medical Association · September 2009

DOI: 10.1001/jama.2009.1365 · Source: PubMed

3 authors:

Jennifer R Kogan
University of Pennsylvania
69 PUBLICATIONS **1,087** CITATIONS
SEE PROFILE

Eric S Holmboe
Accreditation Council for Graduate Medical Edu…
303 PUBLICATIONS **7,776** CITATIONS
SEE PROFILE

Karen E Hauer
University of California, San Francisco
103 PUBLICATIONS **2,083** CITATIONS
SEE PROFILE

Tools for Direct Observation and Assessment of Clinical Skills of Medical Trainees: A Systematic Review

Jennifer R. Kogan; Eric S. Holmboe; Karen E. Hauer

JAMA. 2009;302(12):1316-1326 (doi:10.1001/jama.2009.1365)

REVIEW

Tools for Direct Observation and Assessment of Clinical Skills of Medical Trainees

A Systematic Review

Jennifer R. Kogan, MD
Eric S. Holmboe, MD
Karen E. Hauer, MD

Context Direct observation of medical trainees with actual patients is important for performance-based clinical skills assessment. Multiple tools for direct observation are available, but their characteristics and outcomes have not been compared systematically.

Objectives To identify observation tools used to assess medical trainees' clinical skills with actual patients and to summarize the evidence of their validity and outcomes.

Data Sources Electronic literature search of PubMed, ERIC, CINAHL, and Web of Science for English-language articles published between 1965 and March 2009 and review of references from article bibliographies.

Study Selection Included studies described a tool designed for direct observation of medical trainees' clinical skills with actual patients by educational supervisors. Tools used only in simulated settings or assessing surgical/procedural skills were excluded. Of 10 672 citations, 199 articles were reviewed and 85 met inclusion criteria.

Data Extraction Two authors independently abstracted studies using a modified Best Evidence Medical Education coding form to inform judgment of key psychometric characteristics. Differences were reconciled by consensus.

Results A total of 55 tools were identified. Twenty-one tools were studied with students and 32 with residents or fellows. Two were used across the educational continuum. Most (n=32) were developed for formative assessment. Rater training was described for 26 tools. Only 11 tools had validity evidence based on internal structure and relationship to other variables. Trainee or observer attitudes about the tool were the most commonly measured outcomes. Self-assessed changes in trainee knowledge, skills, or attitudes (n=9) or objectively measured change in knowledge or skills (n=5) were infrequently reported. The strongest validity evidence has been established for the Mini Clinical Evaluation Exercise (Mini-CEX).

Conclusion Although many tools are available for the direct observation of clinical skills, validity evidence and description of educational outcomes are scarce.

JAMA. 2009;302(12):1316-1326 www.jama.com

DIRECT OBSERVATION OF MEDICAL trainees with actual patients by clinical supervisors is critical for teaching and assessing clinical and communication skills. A recent Institute of Medicine report calls for improved supervision of trainees to enhance patient safety and quality of clinical education.[1] The Liaison Committee on Medical Education and Accreditation Council for Graduate Medical Education require ongoing assessment that includes direct observation of trainees' clinical skills.[2,3] By observing and assessing learners with patients and providing feedback, faculty help trainees to acquire and improve skills and help patients through better supervision of clinical care.[4]

Direct observation of medical trainees occurs infrequently and inadequately.[5,6] End-of-rotation global rating forms are often completed by supervisors who have not directly observed trainees with patients.[7] However, assessment based on direct observation should be an essential component of outcomes-based education and certification.[8,9] With current interest in establishing an outcomes-based medical education system that enhances trainee development and patient safety, there is a great need for robust work-based evaluation tools. To our knowledge, a rigorous systematic review has not been performed of the utility and quality of the numerous existing tools for direct observation and assessment of medical trainees with actual patients. We therefore systematically reviewed the literature to determine available tools for direct observation by supervisors of trainees' clinical skills with actual patients. The aim was to describe existing tools and the evidence of their validity and outcomes to provide medical educators with evidence-based assessment mea-

Author Affiliations: Department of Medicine, University of Pennsylvania Health System (Dr Kogan) and American Board of Internal Medicine (Dr Holmboe), Philadelphia, Pennsylvania; and University of California, San Francisco (Dr Hauer).
Corresponding Author: Jennifer R. Kogan, MD, Department of Medicine, University of Pennsylvania Health System, 3701 Market St, Ste 640, Philadelphia, PA 19104 (jennifer.kogan@uphs.upenn.edu).

sures and an understanding of areas for further research.

METHODS

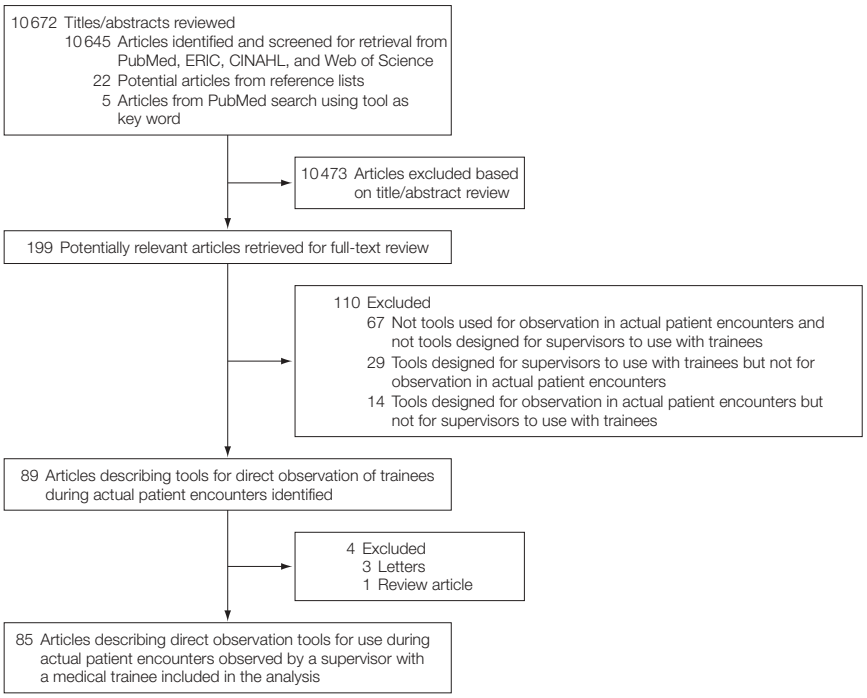
Data Sources

A systematic literature search was conducted using specific eligibility criteria, electronic searching, and hand searching to minimize risk of bias in selecting articles. The search, conducted with the assistance of a library science expert, included relevant English-language studies published between January 1965 and March 2009 using the PubMed, Education Resource Information Center (ERIC), Cumulative Index to Nursing and Allied Health Literature (CINAHL), and Web of Science electronic literature databases. Combinations of terms were used related to competence (*clinical competence*; *clinical skills*), medical education (*education*; *students, education, medical*; *clinical clerkship, internship and residency/methods*; *preceptorship*), and learner level (*student*; *intern*; *resident*). Tables of contents of medical education journals not indexed in PubMed (*Teaching and Learning in Medicine*, 1986-1996; *Medical Teacher*, 1979-1980) were hand-searched. The reference lists of all included articles and identified review articles were examined. A key word search of instruments identified in the included articles was conducted. A more detailed search strategy is available on request.

Study Selection

Studies were included if they described a tool designed (1) for direct observation of skills in clinical settings with actual patients (observer in the room or observing by remote camera) and (2) for use by educational supervisors (interns, residents, fellows, faculty, nurses, nurse practitioners, other trained observers) with medical trainees (medical students, interns, residents, fellows). Studies were excluded that described tools intended (1) for use with standardized patients, (2) for use in simulated settings (eg, without actual patients), or (3) to assess surgical or procedural skills; and (4) without a full article available for review.

Figure. Literature Search and Article Selection Process



Title and Abstract Review

The initial search identified 10 672 citations (FIGURE). All 3 authors independently reviewed citation titles and abstracts to assess eligibility for review, with each title/abstract reviewed by at least 2 authors. Of those, 199 were appropriate for detailed review to determine if they met inclusion criteria. Review articles were excluded. When reviewers disagreed or an abstract was insufficient to determine study eligibility, the full article was retrieved.

Study Review and Data Extraction

A Best Evidence in Medical Education abstraction form[10] was modified to focus on the settings, learners, tool content, and outcomes described in studies. Every article was independently abstracted by 2 authors (J.R.K. and K.E.H.). Each reviewer then reconciled half of the abstractions for completeness and accuracy. Differences in data abstraction were resolved through consensus adjudication. Extracted information included tool characteristics and implementation, validity, and outcomes. Abstracted items characterized tool characteristics (assessed skills, number of items and how they were evaluated, space for open-ended comments or action plan) and implementation (research study design,[11] setting [country, single/multi-institution, specialty, inpatient/outpatient, trainee level], observer characteristics, use for formative/summative evaluation).

Information on reliability and validity was extracted. Although many frameworks to evaluate assessment tools exist,[12-14] the unitary theory of Messick[13] was used. In this approach, validity evidence is used to support the overarching framework of construct validity, the degree to which an assessment measures the underlying construct.[13,15,16] Validity evidence was sought in 5 areas:

- Content: relationship between the tool's content and the construct it intends to measure
- Response process: evidence showing raters have been properly trained (faculty development)

• Internal structure (reliability): internal consistency, test-retest reliability, agreement (interrater reliability), generalizability

Table 1. Characteristics of 85 Studies Describing Tools for Direct Observation of Medical Trainees' Clinical Skills

Characteristics	No. (%)
Location	
United States	59 (69)
Canada	6 (7)
Europe	12 (14)
Australia	4 (5)
Other	4 (5)
Single/multi-institution	
Single institution	64 (75)
Multi-institution	21 (25)
Publication, y	
1970-1979	4 (5)
1980-1989	12 (14)
1990-1999	20 (23)
2000-2009	49 (58)
Setting	
Inpatient	31 (36)
Outpatient	20 (24)
Inpatient and outpatient	20 (24)
Not specified	14 (16)
Specialty	
Emergency medicine	6 (7)
Family medicine/general practice	11 (13)
Internal medicine	40 (47)
Multispecialty[a]	6 (7)
Pediatrics	5 (6)
Psychiatry	2 (2)
Surgery/surgical specialties[b]	9 (11)
Other[c]/not specified	6 (7)
Learners	
Medical students	32 (38)
Residents/fellows	53 (62)
Study design[11]	
Randomized controlled trial	8 (9)
Prospective cohort, historical control, or "pre-post"	8 (9)
Prospective cohort, without baseline	52 (62)
Retrospective cohort	8 (9)
Cross-sectional	3 (4)
Other[d]/not specified[e]	6 (7)
Institutional review board approval	27 (33)
Cost mentioned	11 (13)

[a] Multiple specialties or disciplines included within a single study.
[b] Includes obstetrics/gynecology and ophthalmology.
[c] Includes radiology and anesthesia.
[d] Includes descriptive, qualitative, and survey design. The surveys in this category are surveys of educators about tools, rather than surveys of the observers or trainees who are observed.
[e] Includes studies that did not report a specific statement of study design or articles for which the investigators could not determine the study design.

• Relationship to other variables (concurrent, predictive validity): correlation of scores with other assessments or outcomes; differences in scores by learner subgroups

• Outcomes (educational outcomes): consequences of assessment.

A modified version of Kirkpatrick's hierarchy was used to evaluate outcomes of implementing a tool.[17] Outcome levels abstracted included:

• Participation: learners' or observers' views on the tool or its implementation

• Self-assessed modification of learner or observer attitudes, knowledge, or skills

• Transfer of learning: objectively measured change in learner or observer knowledge or skills

• Results: change in organizational delivery or quality of patient care

Information regarding cost of tool development and implementation was also extracted.[18]

Data Synthesis and Analysis

Due to study heterogeneity, a meta-analysis was not possible. After ascertaining tools used for direct observation, we specifically identified those with evidence of internal structure validity and validity based on relationship to other variables. We determined whether these tools had an educational outcome beyond learners' or observers' attitudes about the tool or its implementation.

RESULTS

Search Results and Article Overview

The Figure summarizes the results of the review process. Of 10 672 citations, 85 met inclusion criteria after title, abstract, and full article review. Fifty-five unique tools were identified. The 85 studies were heterogeneous in their populations, methods, and outcomes (TABLE 1). The most common study design was a prospective cohort without a comparison group. Randomized controlled trials were used in 6 studies in internal medicine,[19-24] 1 in pediatrics,[25] and 1 in an unspecified discipline.[26] Of the studies, 64 (75%) occurred within single institutions. Twenty-seven studies mentioned institutional review board approval.[20-24,27-48] Costs of tool implementation, mentioned infrequently,[37,39,49-57] usually focused on faculty time. One article specifically mentioned administrative costs[56] but none included cost calculations. eTable 1 (available at http://www.jama.com) presents additional information about the characteristics of each study (objective, design, country, learner, specialty, observation location, assessment type [formative/summative], and how observations of trainees occurred).

Description of Tools

Details about each of the tools are provided in TABLE 2. Of the 55 unique tools identified, 21 (38%) were implemented with students, 32 (58%) with residents or fellows, and 2 (the Mini Clinical Evaluation Exercise [Mini-CEX] and 1 unnamed[58]) with both. The largest number of tools (17) were developed or tested in internal medicine settings. The Mini-CEX was the most studied, with adaptations for palliative care,[37] ophthalmology,[59,60] and cardiology[41,61,62] and implementation in multispecialty settings.[63] Most tools contained items on history taking, physical examination, and communication (eTable 2). Eleven tools (20%) contained scales with behavioral anchors.[40,59,60,64-73] Twenty tools (36%) solicited open-ended comments or written action plans. Thirty-two tools (58%) were implemented for formative assessment, 7 (13%) for summative assessment, and 3 (5%) for both, although this distinction was not always clear (eTable 1). Many tools were used once per trainee, although some were used up to 10 times (eTable 1).

Validity Evidence

The frequency of reported validity evidence across tools is summarized in eTable 2. Table 2 describes whether validity was studied for each tool. Actual evidence by study is presented in eTable 3 (eTables are available at http://www.jama.com).

Table 2. Description of Tools (n = 55) Used for Direct Observation of Clinical Skills and the Studies Describing Them

Tool[a]	Specialty	Skills Assessed (Total No. of Items)[b]	Item Evaluation	Validity Evidence[c]				
				Content Validity	Response Process	Internal Structure	Relationships to Other Variables	Outcomes[d]
Tools Used Only With Medical Students								
Amsterdam Attitudes and Communication Scale[74]	Multispecialty	History, communication, counseling, overall (10)	Scale (1-5) with adjective anchors	Yes	Yes	Generalizability coefficient	No	No
Clinical Encounter Card[27,28]	Surgery	History, examination, communication, counseling, overall (8[27]; only 1 of the 8 items evaluated[28])	6-Point normative scale with adjective anchors; open-ended comments	Yes[27]	Yes[27,28]	Interrater reliability; modified generalizability[28]	Concurrent validity; learner level[28]	2[27]
Clinical Skills Assessment Form[88]	Psychiatry	History, examination, communication, counseling (17)	Scale (1-7) with behavioral anchors based on criterion performance	No	No	Interrater reliability; test-retest reliability	No	1, 3
Direct Observation Clinical Encounter Examination[75]	Multispecialty	History, examination, communication, overall (5)	Scale (1-9) with adjective anchors	Yes	Yes	Cronbach α; interrater reliability; generalizability coefficient	Concurrent validity	No
Direct Psychiatric Clinical Examination[89]	Psychiatry	Overall (1)	Scale (7-point) with adjective anchors	No	No	Interrater reliability	Concurrent validity	1, 2
In-training evaluation encounter card[82]	Internal medicine	History, examination, communication, counseling (7)	Scale (1-5) with adjective anchors	No	Yes	Interencounter reliability	No	No
Modified Leicester Assessment Package[64-66]	Multispecialty, family medicine/ general practice	5 categories of consultation competence (multiple)	Numerical scale with behavioral anchors	No	Yes[64-66]	Interrater reliability; generalizability[66]	No	1,[65,66] 4[64]
Murmur learning form[76]	Internal medicine	Cardiac examination (heart murmurs)	Record murmur and whether supervised (yes/no)	Yes	No	No	No	1, 2
Observed long case assessment[67]	Internal medicine	(1)	Behavioral scale; open-ended comments	No	No	No	No	1
Physical examination part I, physical examination part II, interpersonal skills[90]	Internal medicine	History, examination (multiple), communication (30)	Numerical scale; adjective anchors	No	No	Interrater reliability	No	No
Structured Clinical Observation[49]	Pediatrics	History, examination, communication, counseling (52)	Yes/no; open-ended comments for 1-2 items	No	Yes	No	No	1, 3
Structured Single Observer Method[83]	Surgery	Examination (38)	Yes/no	No	Yes	No	Learner level	No
University of Cape Town department of medicine clinical clerkship formative assessment feedback form—bedside presentation[50]	Internal medicine	History, examination, counseling, overall (5)	Scale (1-9) with adjective anchors; open-ended comments	No	Yes	No	No	1, 2
Unnamed[29]	Internal medicine	History, examination, communication, counseling (multiple)	Scale (1-5) with adjective anchors	Yes	Yes	No	No	No
Unnamed[84]	Surgery	Examination, communication (53)	Yes/no	No	Yes	No	Concurrent validity	No
Unnamed[103]	Pediatrics	Examination (multiple)	NR	No	No	No	Learner level	No
Unnamed[51]	Surgery	Technical, interpersonal (18)	Yes/no	No	Yes	No	Learner level	No

(continued)

Table 2. Description of Tools (n = 55) Used for Direct Observation of Clinical Skills and the Studies Describing Them (continued)

Tool[a]	Specialty	Skills Assessed (Total No. of Items)[b]	Item Evaluation	Validity Evidence[c]				
				Content Validity	Response Process	Internal Structure	Relationships to Other Variables	Outcomes[d]
Tools Used Only With Medical Students								
Unnamed[25]	Pediatrics	History, examination, communication (multiple)	Yes/no	No	No	Interrater reliability	No	3
Unnamed[26]	Other/not specified	History, examination (multiple)	Numerical scale	No	No	No	No	3
Unnamed[98]	Internal medicine	History, examination, communication (multiple)	NR	No	No	No	Concurrent validity	No
Unnamed[68]	Other/not specified	History, communication (13)	Scale (1-4) with adjective and behavioral anchors	No	No	Kuder-Richardson 20 reliability coefficients	Concurrent validity	No
Tools Used Only With Residents/ Fellows								
360-Degree evaluation form[30]	Radiology	Communication, counseling (10)	Scale (1-5) with agreement anchors; open-ended comments	Yes	No	Interrater reliability; Cronbach α	Concurrent validity	1, 2
Arizona Clinical Interview Rating Scale (ACIR); History and Physical Exam (HPE) checklist[69]	Multispecialty (family medicine, internal medicine, pediatrics)	ACIR: (14); HPE: history, examination, counseling, communication (58)	Scale (1-5) with behavioral anchors (ACIR); yes/no (HPE)	No	No	Interrater reliability; intercase reliability	Learner level; concurrent validity	No
Clinical Anesthesia System of Evaluation[99]	Anesthesia	Overall (1)	Scale (1-3) with adjective anchors; open-ended comments	No	No	No	Concurrent validity	No
CEX[19,31,32,77,91,100]	Emergency medicine,[32] internal medicine[19,31,77,91,100]	History,[19,31,32,77,91,100] examination,[19,31,32,77,91,100] presentation,[31,32,100] communication,[19,31,77,91] counseling,[19,31,91] diagnosis/plan,[32,100] emergency stabilization,[32] overall[19,31,77] (multiple[32,77,91,100])	Scale (1-9)[32,77]; (1-4)[19,31]; (1-5)[100]; all with adjective anchors; item weighting based on importance[77];yes/no[77]; open-ended comments[19,31,32]	Yes[77]	Yes[19,77]	Accuracy[31]; interrater reliability[19,31,77,91]; item-total correlations[77]; α coefficient[77]; generalizability[77]	Concurrent validity[100]	1,[100] 2[91,100]
CEX; organ system checklists[52]	Internal medicine	History, examination, communication (multiple)	History and communication: scale (1-9) examination: checklist	Yes	No	Interrater reliability; interitem correlations	Concurrent validity	No
Clinical performance biopsy instrument[70]	Family medicine	History/examination, communication, counseling (3)	Scale with behavioral anchors; open-ended comments	No	Yes	No	No	1, 2
Communication behaviors checklist[33]	Emergency medicine	Communication, counseling, overall (34)	Numerical scale with adjective anchors and yes/no	Yes	Yes	Interrater reliability	No	No
Consultation assessment scale[78]	General practice	History, examination, communication, counseling, overall (26)	Scale (1-5) with adjective anchors; open-ended comments	Yes	No	No	No	No
Continuity–Structured Clinical Observations[34]	Pediatrics	History, examination, communication, counseling (46)	Scale (1-3) with adjective anchors; open-ended comments	Yes	Yes	Interrater reliability	No	No

(continued)

Table 2. Description of Tools (n = 55) Used for Direct Observation of Clinical Skills and the Studies Describing Them (continued)

Tool[a]	Specialty	Skills Assessed (Total No. of Items)[b]	Item Evaluation	Validity Evidence[c]				
				Content Validity	Response Process	Internal Structure	Relationships to Other Variables	Out-comes[d]
Tools Used Only With Residents/ Fellows								
Davis Observation Code[79,101]	Family medicine	Disease prevention, health education, health promotion, compliance checking (20)	Yes/No	Yes[79]	No	Interrater reliability[79]	Concurrent validity[79,101]	No
Death Telling Evaluation[35]	Emergency medicine	Counseling (6)	Yes/no; overall (1-3) rating with adjective anchors	No	Yes	No	Learner level	No
Deming management method (adapted)[36]	Emergency medicine	Communication, counseling (16)	Scale (1-3) with adjective anchors	No	Yes	Cronbach α	Concurrent validity	No
Emergency medicine direct observation skills list (3 lists: 1 each for PGY 1, PGY 2, PGY 3)[53]	Emergency medicine	History, examination, communication (40 PGY 1 form; 23 PGY 2 form; 29 PGY 3 form)	Scale (1-5) with adjective anchors; open-ended comments	No	No	No	Concurrent validity	1
First-year resident outpatient core competencies[71]	Family medicine	History, communication, counseling (11)	Behavioral anchors; open-ended comments	No	No	No	No	No
Maastricht History-Taking and Advice Scoring List[72]	General practice	Communication, counseling (11)	Scale (0-6) with adjective and behavioral anchors	No	No	No	Learner level	No
Medical interview skills checklist[54]	Family medicine	History; communication; counseling (83)	Adjective anchors	No	No	No	No	1
Minicard[20]	Internal medicine	History, examination, counseling	Scale (1-4) with adjective anchors; open-ended comments	Yes	Yes	Interrater reliability; alternate forms reliability	No	No
Modified Brown interviewing checklist[104]	Internal medicine	NR	NR	No	No	No	No	1, 2
Ophthalmic Clinical Evaluation Exercise[59,60]	Ophthalmol-ogy	History, examination, communication/ professionalism, counseling (27)	Scale (1-4) with behavioral anchors; open-ended comments	Yes[59]	No	Interrater reliability; Cronbach α[60]	No	No
Palliative care CEX[37]	Internal medicine	Communication, counseling (18)	Yes/no	No	Yes	No	No	1, 2
Patient care–family discussion[21]	Internal medicine	Counseling, self-assessment, overall (30)	Yes/no	Yes	Yes	No	Concurrent validity	1, 2
Patient evaluation assessment form (Michigan State University[38])	Surgery	History, examination, communication, counseling (11)	Scale (0-100) with adjective anchors; open-ended comments	Yes	Yes	No	Learner level	4
Revised infant video questionnaire[39]	Pediatrics	History, examination, communication, counseling (51)	Scale (0-2) with adjective anchors	Yes	Yes	Interrater reliability	"Pre-post"–intervention	3
Standardized Direct Observation Assessment Tool[40]	Emergency medicine	ACGME competencies (26)	3-Point scale with behavioral anchors	Yes	No	Cronbach α; interrater reliability	No	No
Unnamed[92]	Internal medicine	Communication (9)	Yes/no	No	No	Interrater reliability	"Pre-post"–intervention	No
Unnamed[73]	Obstetrics	Knowledge, professionalism, manual skills, overall (4)	Scale (1-7) with behavioral anchors norm referenced for residents' level of training; open-ended comments	No	Yes	Interrater reliability; overall computed G coefficient	Concurrent validity	No

(continued)

Table 2. Description of Tools (n = 55) Used for Direct Observation of Clinical Skills and the Studies Describing Them (continued)

Tool[a]	Specialty	Skills Assessed (Total No. of Items)[b]	Item Evaluation	Validity Evidence[c]				
				Content Validity	Response Process	Internal Structure	Relationships to Other Variables	Outcomes[d]
Tools Used Only With Residents/ Fellows								
Unnamed[55]	General practice	History, communication, counseling, overall (7)	Scale (1-6) with adjective anchors	No	Yes	No	No	1
Unnamed[114]	Internal medicine	History, communication (29)	Scale (1-5) with adjective anchors	No	No	No	No	No
Unnamed[93]	Other/not specified	History, examination, communication, overall (multiple)	Numerical and visual analog scale with adjective anchors	No	No	Interrater reliability; intrarater reliability	"Pre-post"–intervention	1, 3
Unnamed[80]	Internal medicine	History, examination, counseling, overall (10)	Scale (1-5) with adjective anchors	Yes	Yes	Item-total correlation	Learner level; concurrent validity	No
Unnamed[81]	Family medicine	History, communication, counseling (42)	Scale (1-5) with adjective anchors	Yes	No	Interrater reliability	No	No
Unnamed[56]	Family medicine	History, examination (PGY 1) (18); counseling (PGY 2-3) (3)	Competent/not competent (PGY 1); yes/no (PGY 2-3); open-ended comments (PGY 1-3)	Yes	No	No	No	1; 2
Tools Used With Medical Students and Residents/ Fellows								
Mini-CEX[f]	Internal medicine[g]; cardiology[41,61,62]; multispecialty[63,85]; other/not specified[87]	History, examination, communication, counseling, overall,[e] no counseling[41] (7, 6[41])	Scale (1-9)[e] or (1-5)[22,48] or (1-6)[85] with adjective anchors; open-ended comments; overall performance rated on 3-point scale with adjective anchors[47]	No	Yes[h]	Cronbach α[41,42,45,97]; interrater reliability[22,24,63]; interitem correlations[22,42,95,96]; item-total correlations[22,42,45,95,96]; generalizability[22,47,61,96]; reproducibility[63,95]	Concurrent validity[41,42,63,97]; predictive validity[97]; learner level[42,43,52,61,95-97]	1[i]; 2[23,48]; 3[23,48]
Unnamed[58]	Internal medicine	History, examination, communication (153)	Yes/no	No	No	No	Learner level	No

Abbreviations: ACGME, Accreditation Council for Graduate Medical Education; CEX, Clinical Evaluation Exercise; NR, not reported; PGY, postgraduate year.
[a] Tool labeled as unnamed if the tool was not named in the study.
[b] The number of items on the form is greater than the number of skills because the form may have assessed clinical skills in addition to those of interest in this study (data gathering, communication, counseling).
[c] Refers to whether each validity component was studied, not necessarily proven.
[d] Outcomes were rated using a modified Kirkpatrick hierarchy wherein levels of impact were as follows: 1 = participation (learners' or observers' views on the tool or its implementation); 2 = learner or observer self-assessed modification of attitudes, knowledge or skills; 3 = transfer of learning (objectively measured change in learner or observer knowledge or skills); and 4 = results (change in organizational delivery or quality of patient care).
[e] For all citations except ones that follow.
[f] References 22-24, 41-48, 57, 61-63, 85-87, and 95-97.
[g] References 22-24, 42-44, 46-48, 57, 86, and 95-97.
[h] References 23, 41, 42, 44, 45, 47, 48, 62, 81, and 85-87.
[i] References 23, 41, 42, 44, 45, 47, 57, 61, 62, 86, 87, 95, and 96.

Content

Descriptions of tool content selection (content validity) were mentioned for 20 tools (36%)* and typically involved expert or consensus groups reviewing educational competencies and literature.

Response Process

Observers were infrequently trained to use assessment tools. Rater training, described for 47% of tools,† usually occurred once and was brief (10 minutes to 3 hours). Training usually included orienting observers to the tool or discussing feedback principles via e-mail, workshops, or preexisting institutional faculty/resident lectures and meetings.‡ Training sessions that incorporated rater practice using the tool or review of videotaped performances of different competency levels

*References 20, 21, 27, 29, 30, 33, 34, 38-40, 52, 56, 59, 74-81.

†References 19-23, 27-29, 33-39, 41, 42, 44, 45, 47, 49-51, 55, 61, 62, 65, 66, 70, 73-75, 77, 80, 82-87.

‡References 19-22, 27-29, 33-39, 41, 42, 44, 45, 47, 49, 50, 55, 61, 62, 64-66, 70, 75, 77, 80, 82, 85-87.

were described for 8 tools.§ For 2 tools, observers were either given examples of effective feedback[21,85] or trained to provide feedback using role play.[23,49]

Internal Structure

Interrater reliability, reported for 22 tools (40%), was the most commonly reported reliability assessment‖ and was often suboptimal (<0.70).[94] Intrarater reliability[93] and test-retest reliability[88] were reported for 1 tool each. Interitem correlations (correlations between items on the form) and item-total correlations (correlations between items and the overall rating) were reported for 2[22,42,52,95,96] and 4 tools,[22,42,45,77,80,95,96] respectively. Internal consistency was described for only 8 tools¶ but was usually high (Cronbach α approximately ≥0.70).[94] Generalizability/reproducibility coefficients were reported for 8 tools.# Three studies, 1 describing the minicard and the other 2 a modified Mini-CEX, compared performance characteristics of 2 different tools.[20,22,48]

Relationship to Other Variables

Correlation of direct observation scores with other assessments was described for 17 tools (31%) in 22 studies.** Assessments were compared to written examination scores†† and clinical performance ratings.‡‡ Comparisons with objective structured clinical examinations/standardized patient examinations,[28,41,63,73,75,101] chart audits,[79] patient write-ups,[42,68] or patient ratings[30] were less common. In general, correlations were low ($r = 0.1$) or modest ($r = 0.3$).[102] Correlations were disattenuated in 3 studies.[41,73,75]

Performance scores were also compared across training level or other learner characteristics.§§ Eight tools (10 studies) had scores that increased with training level‖‖ ; with 4 tools this trend was not seen.[51,72,83,97] The Mini-CEX had evidence both supporting[24,41,42,61,95,96] and refuting[97] score improvement with training level. With 4 tools, learners' performance improved after clinical skills training and/or feedback.[39,72,92,93]

Outcomes

Surveying trainees and observers about their experiences with a tool was the most common method for assessing outcomes, used with 19 tools (35%).¶¶ Trainees generally rated observation experiences positively.

Modification of trainees' self-assessed knowledge, attitudes, or skills was reported for 9 tools (16%).## Transfer of trainee learning (objectively measured skill or behavior change) was described for 5 tools.[25,26,39,49,93] Studies describing these changes were often nonblinded and failed to control for baseline clinical skills.[26,39]

Outcomes of tool implementation on observer feedback or the effect of observer training on rating behaviors was described for 6 tools.[22,23,27,49,56,70,88] Tool implementation increased the frequency,[27,56] specificity,[70,88] and timeliness[70] of observation and feedback. Training increased confidence using the tool[22,23] but inconsistently improved rater stringency and accuracy.[22,23]

Organizational change was described for 2 tools (Modified Leicester Assessment Package[64]; Patient Evaluation Assessment Form[38]). For both, it was suggested that deficiencies identified on assessments inspired curricular change.[38,64] No tool had evidence that use affected patient care outcomes.

Tools With Multiple Elements of Validity Evidence

Eleven tools had evidence of internal structure validity and validity based on relationships to other variables. These included the Direct Observation Clinical Encounter Examination[75] (multispecialty), Clinical Encounter Card[27,28] (surgery), Direct Psychiatric Clinical Examination[89] (psychiatry), Revised Infant Video Questionnaire[39] (pediatrics), a 360-degree evaluation described by Wood et al[30] (radiology), Davis Observation Code[79,101] (family medicine), Mini-CEX,[41,42,45,47,61,63,95-97] and unnamed tools described by Woolliscroft et al (unspecified discipline),[68] Brennan and Norman[73] (obstetrics), Beckman et al[92] (internal medicine), and Nørgaard et al[80] (internal medicine). Only 3 had evidence of learning. Use of the Revised Infant Video Questionnaire increased learning using a noncontrolled study design.[39] Residents self-assessed improved communication and counseling skills with a 360-degree evaluation.[30] Students reported improved understanding of their history-taking, physical examination, and decision-making skills using the Clinical Encounter Card.[27,28]

COMMENT

Direct observation of medical trainees by faculty remains a vital component of assessment across specialties. Assessment through observation provides ongoing data on trainee performance with actual patients, and effective assessment helps medical educators meet their professional obligation to self-regulate effectively.[105] Enhanced supervision (with observation) can be associated with better patient care and faster acquisition of clinical skills by trainees,[106] and the 2008 Institute of Medicine report recommends greater supervision in medical education to improve patient safety and education.[1] The development of expertise depends on accurate and detailed assessment and feedback.[107] However, faculty and training institutions may not be held accountable for ensuring trainees' clinical competence, and high-quality direct obser-

§References 20, 22, 23, 34, 35, 49, 55, 70, 74, 85.
‖References 19, 20, 22, 24, 25, 28, 30, 31, 33, 34, 39, 40, 52, 60, 63, 66, 69, 73, 75, 77, 79, 81, 88-93.
¶References 30, 36, 40-42, 45, 60, 68, 75, 77, 97.
#References 22, 28, 42, 47, 61, 63, 66, 69, 73-75, 77, 95, 96.
**References 21, 28, 30, 36, 41, 42, 52, 53, 63, 68, 69, 73, 75, 79, 80, 84, 89, 97-101.
††References 21, 28, 41, 42, 73, 75, 84, 89, 97-99, 101.
‡‡References 21, 28, 30, 36, 42, 52, 53, 69, 84, 89, 97, 99-101.
§§References 24, 28, 35, 38, 39, 41, 42, 51, 58, 61, 69, 72, 80, 83, 92, 93, 95-97, 103.
‖‖References 35, 38, 42, 51, 58, 61, 69, 80, 83, 95.
¶¶References 21, 23, 30, 37, 41, 42, 44, 45, 47, 49, 50, 54-57, 61, 62, 65-67, 70, 76, 86-89, 93, 95, 96, 100, 104.
##References 21, 27, 30, 37, 50, 76, 89, 91, 100, 104.

vation of trainees should augment the quality of supervision.[108]

Although we identified many tools available for direct observation of clinical skills, few have been thoroughly evaluated and tested. One tool, the Mini-CEX, has been implemented repeatedly with medical students, residents, and fellows across specialties. The 20 Mini-CEX studies illustrate how validity evidence can accrue and tool implementation can be manipulated (ie, adding behavioral anchors to increase score reliability and accuracy).[20] Multiple publications suggest the validity of Mini-CEX scores. Ten other tools (Table 2) possessing at least 2 levels of validity evidence have potential for wider use with additional research on implementation and consequential validity.*

Although many studies measured trainees' or observers' attitudes about the observation process, few demonstrated improved clinical skills or patient care quality with tool implementation in an educational program. Outcomes such as learning, transfer of skills to new situations, or improved patient care are important and relatively unstudied. Whether these tools are associated with health care system improvements remains an area for future research.

In many studies, rater training (the response process component of validity) was minimally described or did not occur. Whether this omission was related to perceived cost, time constraints, or unawareness of the importance of rater training is unknown. However, observers need training to rate learners' performance reliably and discriminate between performance levels.[8] Randomized trials highlight the value of rater training and its effect on scores.[22,23] Brief training is likely to be ineffective.[19,22,23,77] Although rater training may initially be resource- and time-intensive, these costs should be weighed against potential benefits gained in teaching quality and learning.[18] Given the relative inattention to implementation in the studies we reviewed, as well as the high expense associated with current assessment strategies such as simulation and standardized patient examinations, faculty development that enhances trainees' clinical skills and increases faculty supervision through observation could enhance care and may be cost-effective.

Our findings also suggest several next steps to improve the quality of research in this area. To enhance the quality of evidence in medical education, published research should include the assessment or intervention; methods of implementation; and evidence for reliability, validity, and educational outcomes.[106] However, current research generally does not adhere to these recommendations. After utility of a tool has been demonstrated (validity evidence) and guidelines for implementation developed, randomized study designs should follow whenever possible to assess whether the tool affects educational outcomes.[109,110] More multiinstitutional studies could help improve generalizability of findings. However, these larger, complex studies will require more resources, often lacking for educational research,[111] and may benefit from more streamlined institutional review board approval processes.[112]

A strength of this study is that the review included more than 10 000 abstracts and hand-searching of bibliographies from published studies. However, several limitations should be considered. Publication bias is possible; there are likely tools that have not been described in publications, although they may have relatively poor psychometric characteristics.[113] The search strategy was limited to English-language studies and did not include unpublished abstracts from conference proceedings or nonindexed open-access journals. Although a library science expert assisted with the search, the lack of a specific Medical Subject Heading for direct observation and variability of terms used in the medical education literature may have limited the ability to identify all studies. The literature search may have missed relevant international studies because the search strategy did not include some terms commonly used in non-US countries (eg, registrar).

In conclusion, this systematic review identified and described a large number of tools designed for direct observation of medical trainees' clinical skills with actual patients. Of these, only a few have demonstrated sufficient evidence of validity to warrant more extensive use and testing.

*References 27, 28, 30, 39, 68, 73, 75, 79, 80, 89, 92, 101.

Author Contributions: Dr Kogan had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Kogan, Holmboe, Hauer.
Acquisition of data: Kogan, Hauer.
Analysis and interpretation of data: Kogan, Holmboe, Hauer.
Drafting of the manuscript: Kogan, Holmboe, Hauer.
Critical revision of the manuscript for important intellectual content: Kogan, Holmboe, Hauer.
Statistical analysis: Kogan.
Obtained funding: Kogan, Holmboe, Hauer.
Study supervision: Hauer.
Financial Disclosures: Dr Holmboe reports being employed by the American Board of Internal Medicine and receiving royalties from Mosby-Elsevier for a textbook on physician assessment. No other disclosures were reported.
Previous Presentations: A subset of these data were presented in a poster at the Clerkship Directors in Internal Medicine National Meeting, Orlando, Florida, October 31, 2008.
Funding/Support: This study was funded by a grant from the American Board of Internal Medicine.
Role of the Sponsor: The funding source had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; or preparation, review, or approval of the manuscript.
Additional Information: The 3 eTables are available online at http://www.jama.com.
Additional Contributions: Josephine Tan, MLIS (UCSF) provided help with literature searching; Joanne Batt, BA, and Salina Ng, BA (UCSF), provided administrative assistance and data organization; Patricia S. O'Sullivan, EdD, the ESCape works in progress group (UCSF), and Judy A. Shea, PhD (University of Pennsylvania), provided comments on the manuscript. These individuals did not receive compensation for their roles in the study.

REFERENCES

1. Ulmer C, Wolman DM, Johns MME, eds. *Committee on Optimizing Graduate Medical Trainee (Resident) Hours and Work Schedule to Improve Patient Safety. Resident Duty Hours: Enhancing Sleep, Supervision and Safety*. Washington, DC: National Academy Press; 2008.
2. Accreditation Council for Graduate Medical Education. ACGME Program Requirements for Resident Education in Internal Medicine. http://www.acgme.org/acwebsite/downloads/RRC_progReq/140_internal_medicine_07012009.pdf. Accessed May 29, 2009.
3. Liaison Committee on Medical Education. *Functions and Structure of a Medical School: Standards for Accreditation of Medical Education Programs Leading to the MD Degree*. http://www.lcme.org/functions2008jun.pdf. June 2008. Accessed July 11, 2009.

4. Duffy FD, Gordon GH, Whelan G, et al; Participants in the American Academy on Physician and Patient's Conference on Education and Evaluation of Competence in Communication and Interpersonal Skills. Assessing competence in communication and interpersonal skills: the Kalamazoo II report. *Acad Med*. 2004;79(6):495-507.
5. 2008 AAMC Graduation Questionnaire Program Evaluation Survey: All Schools Summary Report Final. http://www.aamc.org/data/gq/allschoolsreports/2008_pe.pdf. Accessed May 14, 2009.
6. Burdick WP, Schoffstall J. Observation of emergency medicine residents at the bedside: how often does it happen? *Acad Emerg Med*. 1995;2(10):909-913.
7. Epstein RM. Assessment in medical education. *N Engl J Med*. 2007;356(4):387-396.
8. Shumway JM, Harden R; Association for Medical Education in Europe. AMEE Guide No. 25: the assessment of learning outcomes for the competent and reflective physician. *Med Teach*. 2003;25(6):569-584.
9. Crossley J, Humphris G, Jolly B. Assessing health professionals. *Med Educ*. 2002;36(9):800-804.
10. Best Evidence in Medical Education. Appendix IIIA Prototype BEME Coding Sheet. http://www.bemecollaboration.org/beme/files/starting%20reviews/Appendix%20IIIA%20BEME%20Coding%20Sheet.pdf. Accessed May 29, 2009.
11. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. *Designing Clinical Research.* 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2007.
12. Kane M, Crooks T, Cohen A. Validating measures of performance. *Educ Meas Issues Pract*. 1999;18(1):5-17.
13. Messick S. Standards of validity and the validity of standards in performance assessment. *Educ Meas Issues Pract*. 1995;14(1):5-8.
14. van der Vleuten CP, Schuwirth L. Assessing professional competence: from methods to programmes. *Med Educ*. 2005;39(3):309-317.
15. Cook D, Beckman T. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med*. 2006;119(2):166.e7-166.e16.
16. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ*. 2003;37(9):830-837.
17. Kirkpatrick D. Evaluation of Training. In: Craig R, Mittel I, eds. *Training and Development Handbook.* New York, NY: McGraw-Hill; 1967:87-112.
18. van der Vleuten C. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract*. 1996;1(1):41-67.
19. Noel GL, Herbers JE Jr, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents? *Ann Intern Med*. 1992;117(9):757-765.
20. Donato AA, Pangaro L, Smith C, et al. Evaluation of a novel assessment form for observing medical residents: a randomised, controlled trial. *Med Educ*. 2008;42(12):1234-1242.
21. Clay AS, Que L, Petrusa ER, Sebastian M, Govert J. Debriefing in the intensive care unit: a feedback tool to facilitate bedside teaching. *Crit Care Med*. 2007;35(3):738-754.
22. Cook DA, Dupras D, Beckman T, Thomas K, Pankratz V. Effect of rater training on reliability and accuracy of Mini-CEX Scores: a randomized, controlled trial. *J Gen Intern Med*. 2009;24(1):74-79.
23. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. *Ann Intern Med*. 2004;140(11):874-881.
24. Holmboe ES, Huot S, Chung J, Norcini J, Hawkins RE. Construct validity of the Miniclinical Evaluation Exercise (MiniCEX). *Acad Med*. 2003;78(8):826-830.
25. Scheidt PC, Lazoritz S, Ebbeling WL, Figelman AR, Moessner HF, Singer JE. Evaluation of system providing feedback to students on videotaped patient encounters. *J Med Educ*. 1986;61(7):585-590.
26. Stone H, Angevine M, Sivertson S. A model for evaluating the history taking and physical examination skills of medical students. *Med Teach*. 1989;11(1):75-80.
27. Paukert JL, Richards ML, Olney C. An encounter card system for increasing feedback to students. *Am J Surg*. 2002;183(3):300-304.
28. Richards ML, Paukert JL, Downing SM, Bordage G. Reliability and usefulness of clinical encounter cards for a third-year surgical clerkship. *J Surg Res*. 2007;140(1):139-148.
29. Daelmans HE, van der Hem-Stokroos HH, Hoogenboom RJ, Scherpbier AJ, Stehouwer CD, van der Vleuten CP. Feasibility and reliability of an in-training assessment programme in an undergraduate clerkship. *Med Educ*. 2004;38(12):1270-1277.
30. Wood J, Collins J, Burnside ES, et al. Patient, faculty, and self-assessment of radiology resident performance: a 360-degree method of measuring professionalism and interpersonal/communication skills. *Acad Radiol*. 2004;11(8):931-939.
31. Herbers JE Jr, Noel G, Cooper G, Harvey J, Pangaro L, Weaver M. How accurate are faculty evaluations of clinical competence? *J Gen Intern Med*. 1989;4(3):202-208.
32. Shayne P, Heilpern K, Ander D, Palmer-Smith V; Emory University Department of Emergency Medicine Education Committee. Protected clinical teaching time and a bedside clinical evaluation instrument in an emergency medicine training program. *Acad Emerg Med*. 2002;9(11):1342-1349.
33. Rosenzweig S, Brigham TP, Snyder RD, Xu G, McDonald AJ. Assessing emergency medicine resident communication skills using videotaped patient encounters: gaps in inter-rater reliability. *J Emerg Med*. 1999;17(2):355-361.
34. Zimmer K, Solomon B, Siberry G, Serwint J. Continuity-structured clinical observations: assessing the multiple-observer evaluation in a pediatric resident continuity clinic. *Pediatrics*. 2008;121(6):e1633-e1645.
35. Benenson RS, Pollack ML. Evaluation of emergency medicine resident death notification skills by direct observation. *Acad Emerg Med*. 2003;10(3):219-223.
36. Jouriles NJ, Emerman CL, Cydulka RK. Direct observation for assessing emergency medicine core competencies: interpersonal skills. *Acad Emerg Med*. 2002;9(11):1338-1341.
37. Han PK, Keranen LB, Lescisin DA, Arnold RM. The palliative care Clinical Evaluation Exercise (CEX): an experience-based intervention for teaching end-of-life communication skills. *Acad Med*. 2005;80(7):669-676.
38. Anderson CI, Jentz AB, Kareti LR, Harkema JM, Apelgren KN, Slomski CA. Assessing the competencies in general surgery residency training. *Am J Surg*. 2005;189:288-292.
39. McCormick DP, Rassin GM, Stroup-Benham CA, et al. Use of videotaping to evaluate pediatric resident performance of health supervision examinations of infants. *Pediatrics*. 1993;92(1):116-120.
40. Shayne P, Gallahue F, Rinnert S, Anderson CL, Hern G, Katz E; CORD SDOT Study Group. Reliability of a core competency checklist assessment in the emergency department: the Standardized Direct Observation Assessment tool. *Acad Emerg Med*. 2006;13(7):727-732.
41. Hatala R, Ainslie M, Kassen BO, Mackie I, Roberts JM. Assessing the Mini-Clinical Evaluation Exercise in comparison to a national specialty examination. *Med Educ*. 2006;40(10):950-956.
42. Kogan JR, Bellini LM, Shea JA. Feasibility, reliability, and validity of the Mini-Clinical Evaluation Exercise (mCEX) in a medicine core clerkship. *Acad Med*. 2003;78(10)(suppl):S33-S35.
43. Kogan JR, Hauer KE. Brief report: Use of the Mini-Clinical Evaluation Exercise in internal medicine core clerkships. *J Gen Intern Med*. 2006;21(5):501-502.
44. Malhotra S, Hatala R, Courneya C. Internal medicine residents' perceptions of the Mini-Clinical Evaluation Exercise. *Med Teach*. 2008;30(4):414-419.
45. Torre DM, Simpson DE, Elnicki DM, Sebastian JL, Holmboe ES. Feasibility, reliability and user satisfaction with a PDA-based Mini-CEX to evaluate the clinical skills of third-year medical students. *Teach Learn Med*. 2007;19(3):271-277.
46. Holmboe ES, Yepes M, Williams F, Huot S. Feedback and the Mini-Clinical Evaluation Exercise. *J Gen Intern Med*. 2004;19(5 pt 2):558-561.
47. Nair BR, Alexander H, McGrath B, et al. The Mini Clinical Evaluation Exercise (Mini-CEX) for assessing clinical performance of international medical graduates. *Med J Aust*. 2008;189(3):159-161.
48. Cook D, Beckman T. Does scale length matter? a comparison of nine- versus five-point rating scales for the Mini-CEX [published online ahead of print November 26, 2008]. *Adv Health Sci Educ Theory Pract*. doi:10.1007/s10459-008-9147-x.
49. Lane JL, Gottlieb RP. Structured clinical observations: a method to teach clinical skills with limited time and financial resources. *Pediatrics*. 2000;105(4 pt 2):973-977.
50. Burch VC, Seggie JL, Gary NE. Formative assessment promotes learning in undergraduate clinical clerkships. *S Afr Med J*. 2006;96(5):430-433.
51. Reisner E, Dunnington G, Beard J, Witzke D, Fulginiti J, Rappaport W. A model for the assessment of students' physician-patient interaction skills on the surgical clerkship. *Am J Surg*. 1991;162(3):271-273.
52. Woolliscroft JO, Stross JK, Silva J Jr. Clinical competence certification: a critical appraisal. *J Med Educ*. 1984;59(10):799-805.
53. Cydulka RK, Emerman CL, Jouriles NJ. Evaluation of resident performance and intensive bedside teaching during direct observation. *Acad Emerg Med*. 1996;3(4):345-351.
54. Cassata DM, Conroe RM, Clements PW. A program for enhancing medical interviewing using videotape feedback in the family practice residency. *J Fam Pract*. 1977;4(4):673-677.
55. Campbell LM, Howie J, Murray T. Summative assessment: a pilot project in the west of Scotland. *Br J Gen Pract*. 1993;43(375):430-434.
56. Wendling AL. Assessing resident competency in an outpatient setting. *Fam Med*. 2004;36(3):178-184.
57. Hauer KE. Enhancing feedback to students using the Mini-CEX (Clinical Evaluation Exercise). *Acad Med*. 2000;75(5):524.
58. Aloia JF, Jonas E. Skills in history-taking and physical examination. *J Med Educ*. 1976;51(5):410-415.
59. Golnik KC, Goldenhar LM, Gittinger JW Jr, Lustbader JM. The Ophthalmic Clinical Evaluation Exercise (OCEX). *Ophthalmology*. 2004;111(7):1271-1274.
60. Golnik KC, Goldenhar L. The Ophthalmic Clinical Evaluation Exercise: reliability determination. *Ophthalmology*. 2005;112(10):1649-1654.
61. Alves de Lima A, Barrero C, Baratta S, et al. Validity, reliability, feasibility and satisfaction of the Mini-Clinical Evaluation Exercise (Mini-CEX) for cardiology residency training. *Med Teach*. 2007;29(8):785-790.
62. Alves de Lima A, Henquin R, Thierer J, et al. A qualitative study of the impact on learning of the Mini

Clinical Evaluation Exercise in postgraduate training. *Med Teach*. 2005;27(1):46-52.
63. Boulet JR, McKinley DW, Norcini JJ, Whelan GP. Assessing the comparability of standardized patient and physician evaluations of clinical skills. *Adv Health Sci Educ Theory Pract*. 2002;7(2):85-97.
64. Hastings A, McKinley RK, Fraser RC. Strengths and weaknesses in the consultation skills of senior medical students: identification, enhancement and curricular change. *Med Educ*. 2006;40(5):437-443.
65. Hastings AM, Fraser RC, McKinley RK. Student perceptions of a new integrated course in clinical methods for medical undergraduates. *Med Educ*. 2000;34(2):101-107.
66. McKinley RK, Fraser RC, van der Vleuten C, Hastings AM. Formative assessment of the consultation performance of medical students in the setting of general practice using a modified version of the Leicester Assessment Package. *Med Educ*. 2000;34(7):573-579.
67. Newble DI. The observed long-case in clinical assessment. *Med Educ*. 1991;25(5):369-373.
68. Woolliscroft JO, Calhoun JG, Beauchamp C, Wolf FM, Maxim BR. Evaluating the medical history: observation versus write-up review. *J Med Educ*. 1984;59(1):19-23.
69. Swanson DB, Mayewski RJ, Norsen L, Baran G, Mushlin AI. A psychometric study of measures of medical interviewing skills. *Annu Conf Res Med Educ*. 1981;20:3-8.
70. Ross R. A clinical-performance biopsy instrument. *Acad Med*. 2002;77(3):268.
71. Edwards FD, Frey KA. The future of residency education: implementing a competency-based educational model. *Fam Med*. 2007;39(2):116-125.
72. Kramer AW, Dusman H, Tan LH, Jansen JJ, Grol RP, van der Vleuten CP. Acquisition of communication skills in postgraduate training for general practice. *Med Educ*. 2004;38(2):158-167.
73. Brennan BG, Norman G. Use of encounter cards for evaluation of residents in obstetrics. *Acad Med*. 1997;72(10)(suppl 1):S43-S44.
74. de Haes JC, Oort F, Oosterveld P, ten Cate O. Assessment of medical students' communicative behaviour and attitudes: estimating the reliability of the use of the Amsterdam Attitudes and Communication Scale through generalisability coefficients. *Patient Educ Couns*. 2001;45(1):35-42.
75. Hamdy H, Prasad K, Williams R, Salih FA. Reliability and validity of the Direct Observation Clinical Encounter Examination (DOCEE). *Med Educ*. 2003;37(3):205-212.
76. Torre DM, Sebastian JL, Simpson DE. A PDA-based instructional tool to monitor students' cardiac auscultation during a medicine clerkship. *Med Teach*. 2005;27(6):559-561.
77. Kroboth FJ, Hanusa BH, Parker S, et al. The inter-rater reliability and internal consistency of a clinical evaluation exercise. *J Gen Intern Med*. 1992;7(2):174-179.
78. Hays RB. Assessment of general practice consultations: content validity of a rating scale. *Med Educ*. 1990;24(2):110-116.
79. Callahan EJ, Bertakis KD. Development and validation of the Davis Observation Code. *Fam Med*. 1991;23(1):19-24.
80. Nørgaard K, Ringsted C, Dolmans D. Validation of a checklist to assess ward round performance in internal medicine. *Med Educ*. 2004;38(7):700-707.
81. Shapiro J, Schiermer DD. Resident psychosocial performance: a brief report. *Fam Pract*. 1991;8(1):10-13.
82. Hatala R, Norman G. In-training evaluation during an internal medicine clerkship. *Acad Med*. 1999;74(10)(suppl):S118-S120.
83. Dunnington G, Reisner L, Witzke D, Fulginiti J. Structured single-observer methods of evaluation for the assessment of ward performance on the surgical clerkship. *Am J Surg*. 1990;159(4):423-426.
84. Dunnington GL, Wright K, Hoffman K. A pilot experience with competency-based clinical skills assessment in a surgical clerkship. *Am J Surg*. 1994;167(6):604-606.
85. Fernando N, Cleland J, McKenzie H, Cassar K. Identifying the factors that determine feedback given to undergraduate medical students following formative Mini-CEX assessments. *Med Educ*. 2008;42(1):89-95.
86. Kogan JR, Bellini LM, Shea JA. Implementation of the Mini-CEX to evaluate medical students' clinical skills. *Acad Med*. 2002;77(11):1156-1157.
87. Morris A, Hewitt J, Roberts CM. Practical experience of using directly observed procedures, Mini Clinical Evaluation Examinations, and peer observation in pre-registration house officer (FY1) trainees. *Postgrad Med J*. 2006;82(966):285-288.
88. Links PS, Colton T, Norman GR. Evaluating a direct observation exercise in a psychiatric clerkship. *Med Educ*. 1984;18(1):46-51.
89. Price J, Byrne JA. The direct clinical examination: an alternative method for the assessment of clinical psychiatry skills in undergraduate medical students. *Med Educ*. 1994;28(2):120-125.
90. Dawson DJ, Patel VL. Bedside encounter and clinical performance of junior clinical clerks. *Proc Annu Conf Res Med Educ*. 1983;22:186-191.
91. Kroboth FJ, Hanusa BH, Parker SC. Didactic value of the clinical evaluation exercise: missed opportunities. *J Gen Intern Med*. 1996;11(9):551-553.
92. Beckman H, Frankel R, Kihm J, Kulesza G, Geheb M. Measurement and improvement of humanistic skills in first-year trainees. *J Gen Intern Med*. 1990;5(1):42-45.
93. Mir MA, Evans R, Marshall R, Newcombe R, Hayes T. The use of videorecordings of medical postgraduates in improving clinical skills. *Med Educ*. 1989;23(3):276-281.
94. Downing SM. On the reproducibility of assessment data. *Med Educ*. 2004;38(9):1006-1012.
95. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The Mini-CEX (Clinical Evaluation Exercise): a preliminary investigation. *Ann Intern Med*. 1995;123(10):795-799.
96. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The Mini-CEX: a method for assessing clinical skills. *Ann Intern Med*. 2003;138(6):476-481.
97. Durning SJ, Cation LJ, Markert RJ, Pangaro LN. Assessing the reliability and validity of the Mini-Clinical Evaluation Exercise for internal medicine residency training. *Acad Med*. 2002;77(9):900-904.
98. Wiener SL, Koran L, Mitchell P, Schattner G, Fierstein J, Hotchkiss E. Clinical skills: quantitative measurement. *N Y State J Med*. 1976;76(4):610-612.
99. Rhoton MF. A new method to evaluate clinical performance and critical incidents in anaesthesia: quantification of daily comments by teachers. *Med Educ*. 1990;24(3):280-289.
100. Kroboth FJ, Kapoor W, Brown FH, Karpf M, Levey GS. A comparative trial of the Clinical Evaluation Exercise. *Arch Intern Med*. 1985;145(6):1121-1123.
101. Nuovo J, Bertakis KD, Azari R. Assessing resident's knowledge and communication skills using four different evaluation tools. *Med Educ*. 2006;40(7):630-636.
102. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
103. Lagerkvist B, Samuelsson B, Sjolin S. Evaluation of the clinical performance and skill in paediatrics of medical students. *Med Educ*. 1976;10(3):176-178.
104. Roth CS, Schlossberg L, Woods S. Physician-patient communication in ambulatory settings. *Acad Med*. 1996;71(5):558-559.
105. Nasca TJ, Heard JK, Philibert I, Brigham TP, Carlson D. The ACGME: public advocacy before resident advocacy. *Acad Med*. 2009;84(3):293-295.
106. Reed D, Price E, Windish D, et al. Challenges in systematic reviews of educational intervention studies. *Ann Intern Med*. 2005;142(12 pt 2):1080-1089.
107. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Acad Med*. 2004;79(10)(suppl):S70-S81.
108. Durning SJ, Artino AR, Holmboe ES. On regulation and medical education: sociology, learning and accountability. *Acad Med*. 2009;84(5):545-547. doi: 10.1097/ACM.0b013e31819f8031.
109. Beckman TJ, Cook DA. Developing scholarly projects in education: a primer for medical teachers. *Med Teach*. 2007;29(2-3):210-218.
110. Chen FM, Bauchner H, Burstin H. A call for outcomes research in medical education. *Acad Med*. 2004;79(10):955-960.
111. Reed DA, Kern D, Levine R, Wright S. Costs and funding of published medical education research. *JAMA*. 2005;294(9):1052-1057.
112. Dyrbye LN, Thomas M, Mechaber A, et al. Medical education research and IRB review: an analysis and comparison of the IRB review process at six institutions. *Acad Med*. 2007;82(7):654-660.
113. Chaudhry SI, Holmboe E, Beasley B. The state of evaluation in internal medicine residency. *J Gen Intern Med*. 2008;23(7):1010-1015.
114. Meuleman JR, Caranasos GJ. Evaluating the interview performance of internal medicine interns. *Acad Med*. 1989;64(5):277-279.

