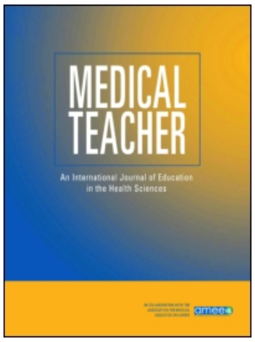
MEDICAL TEACHER
An International Journal of Education in the Health Sciences

Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: https://www.tandfonline.com/loi/imte20

The utility of mini-Clinical Evaluation Exercise in undergraduate and postgraduate medical education: A BEME review: BEME Guide No. 59

Sara Mortaz Hejri, Mohammad Jalili, Rasoul Masoomi, Mandana Shirazi, Saharnaz Nedjat & John Norcini

MEDICAL TEACHER
https://doi.org/10.1080/0142159X.2019.1652732

BEME GUIDE

The utility of mini-Clinical Evaluation Exercise in undergraduate and postgraduate medical education: A BEME review: BEME Guide No. 59

Sara Mortaz Hejri[a], Mohammad Jalili[a,b], Rasoul Masoomi[a], Mandana Shirazi[a,c], Saharnaz Nedjat[d] and John Norcini[e]

[a]Department of Medical Education, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran; [b]Department of Emergency Medicine, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran; [c]Department of Clinical Science and Education at SOS Hospital, Karolina Institute, Stockholm, Sweden; [d]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran; [e]Foundation for Advancement of International Medical Education and Research (FAIMER), Philadelphia, PA, USA

ABSTRACT

Background: This BEME review aims at exploring, analyzing, and synthesizing the evidence considering the utility of the mini-CEX for assessing undergraduate and postgraduate medical trainees, specifically as it relates to reliability, validity, educational impact, acceptability, and cost.

Methods: This registered BEME review applied a systematic search strategy in seven databases to identify studies on validity, reliability, educational impact, acceptability, or cost of the mini-CEX. Data extraction and quality assessment were carried out by two authors. Discrepancies were resolved by a third reviewer. Descriptive synthesis was mainly used to address the review questions. A meta-analysis was performed for Cronbach's alpha.

Results: Fifty-eight papers were included. Only two studies evaluated all five utility criteria. Forty-seven (81%) of the included studies met seven or more of the quality criteria. Cronbach's alpha ranged from 0.58 to 0.97 (weighted mean = 0.90). Reported G coefficients, Standard error of measurement, and confidence interval were diverse and varied based on the number of encounters and the nested or crossed design of the study. The calculated number of encounters needed for a desirable G coefficient also varied greatly. Content coverage was reported satisfactory in several studies. Mini-CEX discriminated between various levels of competency. Factor analyses revealed a single dimension. The six competencies showed high levels of correlation with statistical significance with the overall competence. Moderate to high correlations between mini-CEX scores and other clinical exams were reported. The mini-CEX improved students' performance in other examinations. By providing a framework for structured observation and feedback, the mini-CEX exerts a favorable educational impact. Included studies revealed that feedback was provided in most encounters but its quality was questionable. The completion rates were generally above 50%. Feasibility and high satisfaction were reported.

Conclusion: The mini-CEX has reasonable validity, reliability, and educational impact. Acceptability and feasibility should be interpreted given the required number of encounters.

Background

Assessment plays a central role in medical education. It completes the learning process by monitoring students' progress and achievement regarding the curriculum outcomes. Several tools have been developed for serving this purpose. One of the most frequently-used assessment tools that measure trainees' performance in workplace settings is the mini-Clinical Evaluation Exercise (mini-CEX). An expert, usually a faculty member, observes the actual performance of trainees, rates a variety of their clinical skills, and provides feedback to them (Norcini et al. 1995).

Since its introduction in the 1990s by the American Board of Internal Medicine (ABIM), the mini-CEX has been widely used for different populations and in different contexts around the world. Our scoping search yielded many papers reporting the application of the mini-CEX for either formative or summative purposes. These reports, however, vary in several aspects including the number of required encounters, background of the raters, and the format of the evaluation form. Some of these studies have targeted issues such as psychometric properties, educational consequences, and users' satisfaction with the mini-CEX. In our scoping search, we also found a number of systematic reviews. Some of them included mini-CEX in addition to other workplace-based assessment (WPBA) tools (Kogan et al. 2009; Miller and Archer 2010; Mills et al. 2011; Pelgrim et al. 2011), but four reviews focused solely on the

Practice points

- The mini-CEX can be used in both undergraduate and postgraduate training programs with reasonable validity and reliability.
- Although can be used for summative purposes, by facilitating meaningful feedback and its antecedent favorable educational consequences, the mini-CEX is especially suitable for formative assessment.
- Proper implementation process to ensure psychometric and educational properties while maintaining acceptability and feasibility should be adopted.

CONTACT Mohammad Jalili mjalili@tums.ac.ir Department of Medical Education, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran

Supplemental data for this article can be accessed here.

mini-CEX (Hawkins et al. 2010; Ansari et al. 2013; Sandilands and Zumbo 2014; Lörwald, Lahner, Nouns, et al. 2018). Although each of these reviews focused on one or more characteristics of the mini-CEX, none of them adopted a comprehensive framework to analyze the overall usefulness of the tool.

A common framework to explore all aspects of assessment tools is the utility formula. The framework is composed of validity, reliability, educational impact, acceptability, and the cost of the assessment tool (van der Vleuten 1996). Since this framework provides simple yet comprehensive criteria beyond the traditional psychometrics of validity and reliability, in this systematic review, we evaluated the mini-CEX using van der Vleuten's formula for utility.

Review objective

The main objective of this BEME review was to explore, analyze, and synthesize the evidence considering the utility of the mini-CEX for assessing undergraduate and postgraduate medical trainees. We examined the following factors: reliability, validity, educational impact, acceptability, and cost.

Methods

This systematic review followed the published BEME-approved protocol (Mortaz Hejri et al. 2017) and was in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses), and the STORIES (STructured apprOach to the Reporting In healthcare education of Evidence Synthesis) statements (Gordon and Gibbs 2014).

Search sources and strategies

Our initial scoping search yielded about 500 studies, from which 20 papers were identified as potentially relevant to our review. These findings were used to refine our search strategy. We assumed that including population together with study outcomes would diminish the sensitivity of search strategy and limit the number of retrieved articles even further. Hence, the full search strategy was developed using 11 terms and their Boolean combinations in collaboration with a librarian for health (RM).

We explored seven electronic databases: MEDLINE, EMBASE, PsycINFOvia OVIDsp, CINAHL via EBSCO, ERIC via ProQuest, SCOPUS, and Web of Sciences. The appropriate search strategy for each database was developed individually (Supplementary Appendix 1). The initial search was performed in December 2016 and updated once in June 2017 and again in September 2018. A "tracking tool" was provided as a search log for each database. We stored the search strategies in the database, so we received emails monthly regarding the new updates. We did not limit studies' language or publication year.

To find the gray literature, we searched ProQuest Dissertations & Theses Global. We conducted forward and backward hand searching by checking the reference lists and citations of the included articles and the review articles for additional relevant studies. All citations retrieved were entered into an EndNote database. Duplicate citations were removed.

Study selection criteria

We were interested in studies reporting on the use of mini-CEX in undergraduate or postgraduate medical education which provided empirical data (either quantitative or qualitative) related to the validity, reliability, educational impact, acceptability, or cost of the mini-CEX. No study was excluded on the grounds of study design, geographical location, or language. Yet, commentary and opinion pieces were excluded, as well as review articles, given that studies are required to provide primary data to be included in the BEME review. A summary of the inclusion and exclusion criteria can be found in Table 1. In addition to the original form of the mini-CEX, we included modified versions regardless of whether the competency domains or the rating scale had been changed.

Screening and selection of studies

For screening, all the identified papers were entered into Rayyan (available for free at https://rayyan.qcri.org). Two reviewers (MJ and RM) independently screened the papers in two rounds. The initial screening process was performed based on the titles and abstracts of the papers. In the second round, the full texts of the remaining articles were assessed against the inclusion and exclusion criteria. Studies were included if both reviewers agreed on the relevance. If both reviewers agreed to exclude the paper the article was rejected. In each phase, the percent agreement of raters was calculated by dividing the total number of ratings by the number in agreement (on both included and excluded evidence) to get a fraction, and then converted it to a percentage. In case of any disagreement, the reviewers resolved the issue by discussion and consensus.

Procedure for extracting data

To extract data from the primary studies, we had designed a provisional form (Mortaz Hejri et al. 2017). Before using the form for this review, we revised it by conducting a pilot review for a set of five papers. Some items were modified, and a few other items were added to the form. The data was to be extracted by two independent reviewers (SMH and MS). Before starting extraction, they extracted five articles' data and discussed the results to ensure consistency. Modifications were made when necessary and the form was finalized. Reviewers (SMH and MS), then, started extracting data from all included articles independently. To ensure accuracy, all of the extracted data was checked by a third reviewer (MJ). When there was any discrepancy between the two reviewers, he extracted data himself, and the final decision was made by discussion and consensus.

The data extraction form included the following information: details of the citation, country and institution of study, study aims and design, characteristics of the population, details of the format of the mini-CEX used, methods used to evaluate the utility, and the key findings. Since the number of the included studies were high, where

Table 1. Inclusion and exclusion criteria.

Category	Inclusion criteria	Exclusion criteria
Target population	Undergraduate medical trainee[a] (undergraduate medical education, basic medical education, medical student) or Postgraduate medical trainees[b] (graduate medical education, residency training, resident)	Non-medical students Continuing medical education Continuing professional development Graduate practitioners
Target Intervention	Mini-Clinical Evaluation Exercise in workplace (Mini-CEX, mCEX, Direct Observation of Clinical Skills (DOCS), Clinical Evaluation Exercise (CEX)	Other assessment tools[c] Mini-CEX in non-authentic settings[d]
Outcomes	Utility Validity, credibility Reliability, generalizability, reproducibility, consistency, accuracy Educational impact, educational effect, learning impact, educational outcome, consequential validity Cost, cost-effectiveness, feasibility Acceptability, satisfaction	
Study Language	All languages	
Study type	All designs	No primary empirical research (commentary, opinion pieces, reviews)

[a] Students undertaking undergraduate or basic medical education at a medical school in order to reach a primary qualification in medicine.
[b] Learners of educational programs for medical graduates entering a specialty.
[c] Including studies that reported the mini-CEX besides other assessment tools without presenting the mini-CEX data separately, studies that used tools which differed in several aspects from the original mini-CEX form or focused on a very specific task or domain.
[d] Including studies that used the mini-CEX in simulated settings or in interactions with standardized patients.

Table 2. BEME quality indicators (Buckley et al. 2009).

No	Category	Question
1.	Research question	Is the research question or hypothesis clearly stated?
2.	Study subjects	Is the subject group appropriate for the study being carried out?
3.	Data collection methods	Are the methods used appropriate for the research question and context?
4.	Completeness of data	Attrition rates/acceptable questionnaire response rates
5.	Risk of bias assessment	Is a statement of author positionality and a risk of bias assessment included?
6.	Analysis of results	Are the statistical and other methods of results analysis used appropriate?
7.	Conclusions	Is it clear that the data justify the conclusions drawn?
8.	Reproducibility	Could the study be repeated by other researchers?
9.	Prospective	Is the study prospective?
10.	Ethical issues	Are all ethical issues articulated and managed appropriately?
11.	Triangulation	Were results supported by data from more than one source?

information was not available, authors were not contacted for further information, and it was indicated as "not reported."

Study quality assessment

To evaluate the methodological quality of the eligible studies, we used a generic checklist, developed by Buckley et al. (Buckley et al. 2009), which is applicable to all study designs. The checklist consists of 11 criteria, each is rated as "met," "unmet," or "unclear" (Table 2). To be deemed of high quality, studies are required to meet a minimum of seven indicators. Although the high-quality papers were evaluated in greater depth, we did not remove any study because of poor quality.

The quality assessment was done independently by two team members (SMH and MS).

A third reviewer (MJ) checked the marks, and when there were any disagreements between the two reviewers, he evaluated the paper himself, and then the issue was resolved through discussion and consensus.

Synthesis of extracted evidence

First, we provided a description on the characteristics, setting, and context of the included studies. This descriptive synthesis was used as the basis of synthesis evidence to address the review questions and objectives. In attempting to answer our original research questions, we presented our findings according to five outcomes (validity, reliability, acceptability, educational impact, and cost). Considering the reliability of the mini-CEX, we anticipated to find mostly quantitative data. It was expected to have both qualitative and quantitative data for questions addressing the validity and acceptability of the tool. On the other hand, educational impact was supposed to have been dealt with mainly through qualitative methods.

With regard to quantitative data, we predicted significant heterogeneity in studies' setting, design, and methodology that would preclude meta-analysis for most items. Hence, we planned to report our findings narratively and undertook a rich and exploratory descriptive synthesis of evidence to explain differences in findings. Just for Cronbach's alpha, we managed to conduct a meta-analysis by pooling data using a random-effects method considering study sample size and number of items. A normalizing transformation for F statistic was applied in addition to a weighting scheme involving both the estimated sampling error of each study and the estimated random effects variance (Rodriguez and Maeda 2006).

Results

Search results

Figure 1 illustrates the process of literature searching and selection.

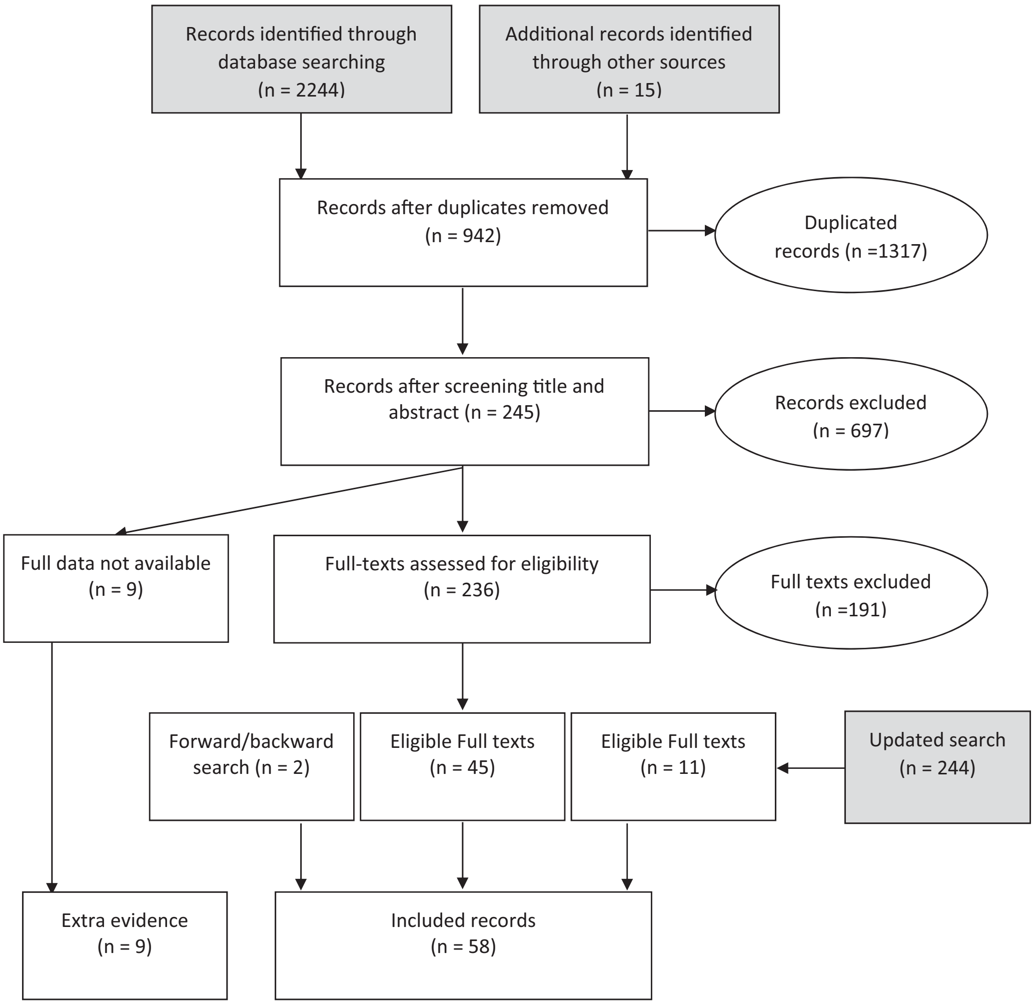
Records identified through database searching (n = 2244)
Additional records identified through other sources (n = 15)
Records after duplicates removed (n = 942)
Duplicated records (n =1317)
Records after screening title and abstract (n = 245)
Records excluded (n = 697)
Full data not available (n = 9)
Full-texts assessed for eligibility (n = 236)
Full texts excluded (n =191)
Forward/backward search (n = 2)
Eligible Full texts (n = 45)
Eligible Full texts (n = 11)
Updated search (n = 244)
Extra evidence (n = 9)
Included records (n = 58)

Figure 1. Flow diagram of study selection process.

We retrieved a total of 2259 citations; 2244 of them were obtained from electronic databases and 15 abstracts were yielded as the result of searching the gray literature. All of them were imported into EndNote. After removing 1317 duplicates, a total of 942 abstracts remained.

In the first round of screening, 637 and 145 papers were excluded and included by both reviewers, respectively. The percent agreement between raters at this stage was 0.83. All 160 disagreements were resolved by discussion, which resulted in a further 60 papers being excluded. Consequently, 697 papers were excluded. The reasons included review articles, studies without data relevant to study questions or application of mini-CEX for health professions other than medicine.

Nine citations were conference abstracts and Really Good Stuff, so their full data were not available. According to the Cochran handbook, a considerable number of conference abstracts never get published in journals, and those that are finally published in full, demonstrate totally different findings (Higgins and Green 2011). Hence, we did not exclude those nine studies but treated them separately for extracting their data as extra evidence.

In the second round of screening, the percent agreement was 96.8%, as both reviewers agreed on inclusion/exclusion of 185 papers. Again, all disagreements were handled by discussion. Tools that have been devised based on the mini-CEX but were totally different in terms of their format and purpose, were excluded at this stage. Some examples include: P-MEX (Cruess et al. 2006), Ophthalmic Clinical Evaluation Exercise (OCEX) (Golnik et al. 2004), Palliative Care CEX (Han et al. 2005), eCEX (Ferenchick et al. 2010), Resident Observation and Competency Assessment (RO&CA) (Musick et al. 2010), Clinicosocial Case Study (Gohel et al. 2016), and Practical physiology (Gade et al. 2017). Moreover, papers which used mini-CEX either in simulated settings such as Objective Structure Clinical Exams (OSCEs), or for evaluating raters' performance in video-taped clinical encounters (Holmboe et al. 2003; Margolis et al. 2006; Donato et al. 2008; Cook et al. 2009; Cook and Beckman 2009; Lie et al. 2010; Alves de Lima et al. 2013; Yeates et al. 2013; Gingerich et al. 2017) were also excluded.

We updated the search in June 2017 and September 2018, which resulted in 150 and 94 more citations, respectively; from which 11 papers were considered as relevant. Hence, 58 papers were ultimately included for data extraction.

Overview of the included studies

Supplementary Appendix 2 reports details of all included studies.

Regarding study design, all papers were either description or justification studies according to the classification suggested by Cook et al. (2008). Characteristics of the included studies are presented in Table 3.

Among the included studies, several interesting clustering of papers was noticed. These involved a total of 26 studies, which might affect the conclusions due to the

Table 3. Characteristics of the included studies.

Characteristic	Categories	Number of studies
Design	Qualitative	7
	Quasi experimental	3
	Mixed method	3
Study site	USA or Canada	16
	Europe	12
	Other (Australia, Argentina, or Asia)	30
Language	English	51
	Spanish	4
	Danish	1
	Turkish	1
	French	1
Program	Postgraduate[a]	33
	Undergraduate	21
	Foundation Year	3
	IMGs[b]	1
Setting	Outpatient	5[¥]
	Inpatient	4[£]
	ED	4[€]
	OR	1[µ]
	Mixed settings	25
	Not stated	19

Characteristic	Min	Max	Median
Duration of study[c]	2 months (Alves De Lima et al. 2005)	5 years (Pernar et al. 2011)	
No. of participants	8 (Ergin et al. 2013; Fernández Galvez 2011)	1149 (Weller et al. 2017)	80
No. of raters	4 (Alves De Lima 2005)	2401 (Weller et al. 2017)	67
No of forms completed	16 (Alves De Lima 2005)	7808 (Weller et al. 2017)	365

[a]Eleven of which implemented the mini-CEX in more than one program or hospital, and the rest were single-center.
[b]International Medical Graduates.
[c]Duration of study was ≥12 months in 31 reports, and <1 year in 12 studies.

potential correlation of data. Five studies were conducted in anesthesiology residency in New Zealand and Australia (Weller, Jolly, et al. 2009; Weller, Jones, et al. 2009; Weller et al. 2014; Castanelli et al. 2016; Weller et al. 2017). Three studies were published using data of cardiovascular residents in Argentina (Alves de Lima et al. 2005; Alves De Lima et al. 2007; Alves de Lima et al. 2010). Another three reports were from implementing mini-CEX within the residency programs of the US from the same investigator (Norcini et al. 1995; 1997; 2003). A further three citations were related to 1-month rotation of residents of different departments at emergency departments in Taiwan (Lin et al. 2012; Chang et al. 2013, 2017). Another three studies were performed in pediatrics residency in India (Singh and Sharma 2010; Goel and Singh 2015; Gupta et al. 2017). Among studies on undergraduate programs, four in the US (Torre et al. 2007; Torre et al. 2011; Kogan et al. 2003; Ney et al. 2009), three in Switzerland (Montagne et al. 2014; Rogausch et al. 2015; Berendonk et al. 2018), and two in the UK (Hill and Kendall 2007; Hill et al. 2009) were conducted by the same research teams.

The mini-CEX format and process in the included studies

The original format of mini-CEX, as it appeared in the very first articles (Norcini et al. 1995; 1997; Durning et al. 2002), was composed of four competency domains; namely, history taking, physical examination, clinical judgment and synthesis, and humanistic qualities (Supplementary Appendix 2). The ABIM later revised the domains, so the final form included medical interviewing, physical examination, humanistic qualities/professionalism, clinical judgment, counseling, and organization/efficiency. Most of the subsequent studies (35 studies) have applied the 6-domain form, without any changes or with minor changes in wording or layout. However, some investigators modified the mini-CEX domains by removing or adding a number of competencies. These modifications have resulted in forms with different number of domains, depending on the competencies of interest: five (Hatala et al. 2006; Ney et al. 2009), seven (Lin et al. 2012; Chang et al. 2013, 2017), eight (Suhoyo et al. 2014), nine (Weller, Jolly, et al. 2009; Weller, Jones, et al. 2009), 10 (Weller et al. 2014, 2017), or 13 (Castanelli et al. 2016). The extra domains include areas such as technical and procedural skills, patient education, risk management, documentation, and team interaction. All of the above-mentioned forms included an overall rating of trainees' performance, except the forms used in three studies that conducted in Taiwan (Lin et al. 2012; Chang et al. 2013, 2017).

In accordance with the original format of the mini-CEX, most studies (45 out of 53 studies that reported their scaling format) used a 9-point Likert scale for rating each competency domain in which 1, 2, and 3 were defined as unsatisfactory, 4 as marginal, 5 and 6 as satisfactory, and 7, 8, and 9 as superior performances. Yet, a variety of rating scales in terms of lengths and labels can be seen in the included studies. These modifications consist of 4-point (Suhoyo et al. 2014), 5-point (Cook et al. 2010), 6-point (Hill and Kendall 2007; Hill et al. 2009), 8-point (Ergin et al. 2013; Playford et al. 2013), and 10-point (Montagne et al. 2014; Rogausch et al. 2015) rating scales. In all but nine studies, the rating form also contained a "not observed" or "insufficient contact" option.

Weller and colleagues in several studies changed the scoring system from the conventional scoring system to an entrustment scale. This system was based on the trainee's level of independence with a case and the raters scored trainees according to the level of supervision requirement (SReq). The investigators corrected the raw SReq scores for case difficulty and also calculated the observed minus expected (O-E) SReq scores (Weller et al. 2014, 2017).

Eleven studies reported the use of a digital tool to facilitate the implementation of mini-CEX in their settings. These include online and computer-based format in most cases (Weller, Jolly, et al. 2009; Weller, Jones, et al. 2009; Cook et al. 2010; Playford et al. 2013; Weller et al. 2014; Castanelli et al. 2016; Weller et al. 2017). Torre et al. deployed a PDA-based form (Torre et al. 2007, 2011). Chang and colleagues described that in the first few months, the assessments were paper-based; however, over the following months, part of the process changed to a computer-based format (Chang et al. 2013; 2017).

Out of 41 papers which explicitly stated their purpose of assessment, 33 declared that the gathered data were not meant to contribute to final course grades and were used only for formative purposes. Seven citations described both formative and summative usage (Hill and Kendall 2007; Weller, Jones, et al. 2009; Brazil et al. 2012; Playford et al. 2013; Suhoyo et al. 2014; Weston and Smith 2014; Yanting et al. 2016). In one of their articles, Alves de Lima and colleagues mentioned that they first introduced the mini-CEX as a formative assessment, and then after 2 years, started to use it as a summative tool (Alves de Lima et al. 2010).

Twenty-nine papers stated a minimum number of required mini-CEX encounters, which varied from 3 to 23. However, these numbers should be adopted by the duration of educational program. Twenty-two out of these 29 papers reported the duration, which ranged from 1 month (two studies) to 1 year (six studies). We calculated the average numbers per month to standardize the records. The stringent requirement was related to Taiwanese residents from different disciplines, required to complete four mini-CEXs, while they were spending a 1-month rotation in emergency department (Lin et al. 2012; Chang et al. 2013). On the other hand, in a cardiology residency program in Argentina, and a pediatrics residency in India, each resident was supposed to fulfill about half assessment per month that is 4–6 times in a year (Alves de Lima et al. 2010; Goel and Singh 2015).

While the usual procedure of mini-CEX is trainee-initiated, a few studies mentioned that the assessments were performed at pre-planned times (i.e., they were scheduled for certain dates in advance) (Alves de Lima et al. 2005; Hatala et al. 2006; Singh and Sharma 2010; Khalil et al. 2017). Moreover, some studies stated that the person responsible for selecting the cases was the rater, not the student (Fernando et al. 2008; Lin et al. 2012; Suhoyo et al. 2014; Chang et al. 2017). In three studies, each trainee was evaluated by more than one rater on each occasion (Abadie et al. 2015; Khalil et al. 2017; Pottier et al. 2018).

Two studies declared that no training session was held for the raters (Weston and Smith 2014; Pottier et al. 2018). However, in 40 papers some kind of training, for raters and/or trainees, was mentioned. These included written and online guidelines, departmental presentations, interactive workshops, video-based materials, etc.

Methodological quality of the included studies

Forty-seven (81%) of the included studies met seven or more of the quality criteria. The scores ranged from 5 (Playford et al. 2013; Saeed et al. 2015; Meresh et al. 2018) to 11 (Kogan et al. 2003; Holmboe et al. 2004; Nair et al. 2008; Kim et al. 2016; Humphrey-Murto et al. 2018).

While study designs and data collection methods were generally aligned with the stated research question, 10 papers failed to clearly state their research question or hypothesis. Lack of obtaining an ethical approval was a limitation in 20 studies. In general, triangulation was found in just 20 studies. Twenty-seven studies explicitly stated their risk-of-bias inherent to the study design or methodology.

Score of each included study has been indicated in Supplementary Appendix 2.

Overview of the utility in the included studies

Considering the desired outcomes of implementing mini-CEX, we were looking for data on five criteria of utility; namely, validity, reliability, educational impact, acceptability, and cost. However, this was by no means a straightforward task and we came up with papers reporting a variety of outcomes which made it difficult to synthesize the extracted data. Prior to synthesis, we had to discuss the possible options to decide on a unified procedure for the purpose of this review. We divide studies into the following categories, in this regard:

- The first group of studies not only used the exact utility indicators that we had previously defined for this review but also provided a clear description of the methodology used to achieve that outcome. For example, some studies stated that for "reliability," "Cronbach alpha coefficient" was calculated. We imported these studies' data as they were.
- The second group of papers reported an outcome with a name which was not in our list. However, by taking a closer look, we could relate the reported outcomes with one of the outcomes we had in mind. For instance, a considerable number of papers gathered data regarding satisfaction of raters and students. We had little trouble reaching a consensus for this type. We simply put "satisfaction" under "acceptability." Likewise, a large number of papers reported the mini-CEX feasibility. Since there is no such term in the original utility formula, and since almost none of the studies had performed the cost analysis, we decided to put them under "cost." Then for more clarity, we changed "cost" to "cost and feasibility."
- The third group of papers caused serious bother, as they reported only the methods of data gathering, without mentioning a particular outcome. For example, several studies investigated the feedback provided to the students, in terms of their types, formats, impacts on learning, etc. Having discussed the issue in depth and in detail, we categorized these findings into "educational impact." Similarly, we considered completion rate, observation time, and feedback time as the "feasibility" evidence. In addition, almost all studies reported the correlation between domains (such as history taking and physical exam). We concluded that these correlations demonstrate some evidence of "construct validity."

Table 4. Methods of data gathering for assessment of the five utility criteria in the included studies and the frequency of each method.

Utility Index (number of studies)	Method of data gathering	Number of studies
Reliability (28)	• Cronbach's alpha coefficient (internal consistency)	18
	• Generalizability analysis: G and D studies	11
	• Standard error of measurement and confidence interval	8
	• Inter-rater agreement	4
	• Inter-encounter reliability	1
	• Perceptions of stakeholders through survey, focus group discussions, or interviews	2
Validity (33)	• Representativeness of observed patients' problems and diagnoses (content validity)	3
	• Measuring the discrimination between different competency levels (construct validity)	18
	• Pearson's coefficient or regression analysis among components of the competence and among 6 competencies and the overall competence (construct validity)	10
	• Factor analysis (construct validity)	4
	• Correlation between the mini-CEX scores and the results of other examinations (criterion validity)	9
	• Perceptions of stakeholders through survey, focus group discussions, or interviews	4
Acceptability (36)	• Satisfaction of the assessors and trainees via questionnaires	29
	• Satisfaction of the assessors and trainees via interviews, or focus group discussions	8
Educational impact (30)	• Assessing the change in performance of trainees	3
	• Asking trainees' and faculty's opinions about the utility of the mini-CEX as an educational tool (interviews, surveys, and focus group discussions)	20
	• Frequency of feedback provision, and analysis of feedback content	8
Cost and feasibility (38)	• Transforming the total annual faculty time expenditure into dollars	1
	• Time spent per encounter (observation, feedback, total)	29
	• Completion rate	21
	• Percentage of incomplete data in the mini-CEX forms	1
	• Perceptions and experiences of stakeholders (through surveys, individual or group interviews)	7

- In a few cases, disagreement was seen among studies when they were addressing a certain method. For instance, one paper compared the mini-CEX scores during the course to see if any progress had happened, and considered this as validity, while another one claimed this is the educational impact of the tool. We considered this as validity, since all improvement during a period of time cannot be attributed to the assessment.

In this way, we managed to organize all of the extracted data in a meaningful and consistent approach. We found out that, in five studies out of 58 citations, just one utility criterion had been assessed, and only two studies evaluated all five outcomes. Twenty-two, 11, and 18 studies investigated two, three, and four criteria, respectively (Supplementary Appendix 2).

Reliability

Twenty-eight studies reported data on the reliability (Table 4).

Internal consistency

Eighteen values for Cronbach's alpha were taken over competency domains in the mini-CEX ranging from 0.58 (Paravicini and Peterson 2015) to 0.97 (Fernández Galvez 2011). More than half of the studies (11 out of 18) calculated a coefficient of around or higher than 0.9. While an acceptable value is considered above 0.7, two studies calculated a coefficient of less than 0.7 (0.64 in Eriksen et al. 2009 and 0.58 in Paravicini and Peterson 2015). We could not find any special characteristics in these studies except one. This study on 23 residents in a chiropractic program included three sessions over a period of 12 months. Each session contained four stations with separate encounters with real patients. The focus of stations was focused history taking followed by clinical reasoning, physical examination followed by clinical reasoning questions, patient management after history taking, and a radiology station consisting of 5 different imaging cases. They used a 9-point scale for two first sessions and a 5-point scale for the last session (Paravicini and Peterson 2015). Meta-analysis was performed using data from 16 studies which had reported sufficient information. Two studies were not enrolled for meta-analysis since the number of their participants had not been reported. The weighted mean for Cronbach's alpha within a random effect model was found to be 0.90 (95% CI: 0.89–0.91).

Generalizability analysis

Generalizability analysis, which quantifies errors of measurement form different sources in a test, based on analysis of variance (ANOVA), was performed in 11 studies. Coefficients varied noticeably as a function of number of encounters, when using D study to estimate the reliability of varying combinations of numbers of assessors and cases per trainee, the results are diverse. Norcini et al. reported a G coefficient of 0.55 for four observations (Norcini et al. 1995). Alves de Lima and colleagues estimated that for one encounter G coefficient would be 0.07 and for 50 encounters it would reach to 0.80 (Alves de Lima et al. 2007). In another G analysis considering eight encounters, the G coefficient was 0.88 (Nair et al. 2008). One study calculated G coefficient for 3–12 encounters and found it to be between 0.60 and 0.86 (Jackson and Wall 2010). Hill and colleagues calculated G coefficients separately within clinical rotations and across clinical rotations: With regard to just one encounter, G was 0.18 in both data sets, however when 15 encounters were taken into account, G coefficient was 0.77 and 0.73, within and across clinical rotations, respectively (Hill et al. 2009). One study performed G study regarding two different designs: When one assessor-one case was considered, G coefficient was 0.07 and 0.04 in fully crossed design and fully nested design, respectively.

Using 25 assessors-10 cases (250 encounters) would result in G coefficient of 0.84 in fully nested design, while for 15 assessor-5 cases (75 encounters), G coefficient was estimated as 0.4 in a fully nested design (Weller, Jolly, et al. 2009). Cook et al. applied a 5-point Likert scale, but they adjusted their analysis for the original 9-point scale. According to their findings, for one assessor and 14 raters, reliability coefficient was 0.19 and 0.75, respectively (Cook et al. 2010). A recent study by Humphrey-Murto and colleagues used the G study by including (separately) the discipline and rater type as a facet and came up with G coefficients of 0.53 and 0.23, respectively (Humphrey-Murto et al. 2018).

Generally, if looking at the number of encounters needed to achieve a reliability coefficient of 0.7, different numbers of encounters were estimated: five (Jackson and Wall 2010), 10–14 (Cook et al. 2010), and 11–13 (Hill et al. 2009) encounters have been suggested in different studies. However, in some studies, a higher number of mini-CEXs were needed to achieve this goal. One study reported a minimum number of 30 encounters (Alves de Lima et al. 2007). Weller et al. stated that at least 60 encounters (20 assessors each observing three cases) would be needed in a fully crossed design; they failed to reach a desirable G coefficient in nested design, regardless of the number of raters or cases (Weller, Jolly, et al. 2009). These investigators in another study showed that the composite scoring system requires eight encounters (eight raters each assessing one case) to reach a reliability of 0.7. Assessor numbers could be reduced to seven if each assessed three cases. This moderate level of reliability cannot be obtained if the overall score is taken into account even with 50 encounters (10 assessors × 5 cases) (Weller et al. 2014). Weller et al. believed that their innovative scoring system was more intuitive for raters and expected it to improve assessment accuracy. In fact, the findings of their studies showed that a moderate level of reliability (G coefficient >0.7) can be achieved with as low as 9 encounters (3 assessors with 3 cases each) when corrected independence scores are considered (Weller et al. 2014). D study results showed that considerably fewer assessments (4 assessors each rating 2 cases or 6 assessors each rating one case) can result in moderate reliability if O-E SReq scores are used (Weller et al. 2017).

In some papers, variance components of different factors were reported as well as the G coefficients. In one study only 6% of total variance was related to the desirable variance (resident variance) and near 75% was associated with differences between assessors/cases and residual error (Alves de Lima et al. 2007). Cook et al. found that the object of measurement (the resident) accounted for only about 12% of the variance, which was less than the variance arising from other sources (Cook et al. 2010). In another study, 40% of score variance was found to be due to assessor stringency (Weller, Jolly, et al. 2009). Another study by the same investigator looked at the contribution of trainee ability, assessor stringency, assessor subjectivity (across trainees), and residual case-to-case variation in scores variations. The study concluded that composite scores reflect trainee ability better than the overall scores. When composite scores of the domains are considered, the variance due to trainee ability, assessor stringency, assessor trainee-related subjectivity, and residual case-to-case variation were 22%, 40%, 22%, and 16%, respectively. For overall scores, the variance components were 9% for trainee ability, 29% for assessor stringency, 26% for assessor trainee-related subjectivity, and 36% for residual case-to-case variation (Weller et al. 2014). In a recent study, error constituted the majority (66–85%) of the variation in scores and students accounted for only a small percentage (3–6%) of the variance (Humphrey-Murto et al. 2018). Having applied a scoring system other than the conventional method, Weller and colleagues concluded that using O-E SReq scores resulted in favorably large variance component pertaining to trainees' ability (Weller et al. 2017).

Standard error of measurement (SEM) and confidence interval (CI)

CI and SEM are used to estimate the amount of error in any measurement. Values for SEM/CI were obtained from eight studies. Depending on the number of encounters and the design of the study (i.e., crossed vs. nested) and whether the SEM is calculated for overall score or average score, the magnitude of SEM varied greatly. In one study, SEM was calculated to be 0.35 for four encounters (Norcini et al. 1995), while another study found the same value of SEM for eight encounters (Nair et al. 2008). SEM was reported 0.52–0.53 for one encounter; and 0.13–0.15 when 15 encounters (Hill et al. 2009). Alves de Lima et al. showed that for one and 50 encounters, SEM was 0.78 and 0.11, respectively (Alves de Lima et al. 2007). Weller et al. found that in a fully crossed design, 250 encounters (25 assessors ×10 cases) would result in a CI of 0.17 while a fully nested design would give rise to a CI of 0.40 with 75 (15 × 5) encounters (Weller, Jolly, et al. 2009). Cook et al. also noted a SEM of 0.26–0.29 for 14 encounters (Cook et al. 2010).

An acceptable level of SEM (<0.26) or CI (<0.52) would be achieved by a minimum of four encounters (Hill et al. 2009), 10 encounters (Alves de Lima et al. 2007), 10 assessor-one case (10 encounters) (Weller, Jolly, et al. 2009).

Other quantitative methods

Hatala and colleagues, considering the fact that each encounter was observed by a separate examiner, used Cronbach's alpha to examine the inter-encounter reliability of the candidates' mean overall score. The coefficient was calculated to be 0.74 (Hatala et al. 2006). In one study, the investigators performed a one-way ANOVA of one factor between the scores of global competence of each of the teachers to analyze the inter-observer variability. They noticed that the scores of global competence of each of the teachers was significantly ($p < 0.0001$) different (Fernández Galvez 2011). In another study in two institutions in Argentina, however, the findings were different. Using the Bland Altman method for determining inter-observer agreement, researchers found good concordance in the assignment of scores between observers in both hospitals (limit of agreement in hospital 1 was −1.952 to 1.507 and limit of agreement in hospital 2 was −2.148 to 1.920). They concluded that the "values of disagreement did not exceed figures of operational relevance" (Abadie et al. 2015). In another recent study in Switzerland, in which each encounter was observed by two assessors, the researchers analyzed the inter-rater agreement on categorical data (such as unsatisfactory, satisfactory, superior, etc.) using Cohen's kappa coefficient and found that the coefficient for all categories using the 9-point evaluation scale was 0.31 (fair) and for the 5-point scale was 0.42 (moderate). The overall percent agreements for the 9- and 5-point scales were 43% and 55% respectively. The authors also assessed the inter-rater reliability of the actual numerical values for the various competencies at a station using the intra-class correlation coefficient (ICC) and found that ICC values ranged between 0.39 and 0.77 for the 9-point scale, and between 0.14 and 0.70 for the 5-point scale (Paravicini and Peterson 2015). Another study in France yielded similar results for ICC, namely between 0.39 and 0.81 (Pottier et al. 2018). Hill and colleague, using an exploratory variance component analysis of domain scores to examine the contribution of relevant factors to students' scores, revealed that examiner stringency contributed significant unwanted variation to the scores with values as great as 23% per encounter and 29% per competency domain (Hill et al. 2009).

Qualitative methods

When asked about their perceptions about the mini-CEX, both examiners and students agreed that several short assessments would be more reliable than one long case (Hill et al. 2009). In a study in anesthesiology, some faculty raised the issue that there is potential for personal interactions leading to strong bias and some stated that the face to face nature of the encounter encouraged leniency and above-average scores (Weller, Jones, et al. 2009).

Validity

A total of 33 studies reported data on different aspects of the mini-CEX validity. The investigators have applied several methods to gather validity evidence (Table 4).

Representativeness of the performed mini-CEXs (content validity)

Three studies reported good content coverage of the cases and observations. One study conducted in five internal medicine residency program sites concluded that the observations "covered a broad range and included a representative array of common problems" (Norcini et al. 1995). Hatala and colleagues, also, showed that trainees had been observed across a broad range of common problems and diagnoses. They stated that these cases were representative of the domain of internal medicine (Hatala et al. 2006). Another study showed that the examinations covered a broad range of problems in cardiology (Alves de Lima et al. 2007).

Discrimination between different competency levels (construct validity)

Some studies compared the performance of the learners with varying preexisting levels of competency, such as residents in different years of training (Norcini et al. 1995; Alves de Lima et al. 2007; Jackson and Wall 2010; Fernández Galvez 2011; Liao et al. 2013; Weller et al. 2017), while others tracked the change of ratings over the course of time, e.g., in consecutive academic quarters, blocks, or even months of training to see the progression of ratings over the span of the study period (Kogan et al. 2003; Norcini et al. 2003; Playford et al. 2013; Suhoyo et al. 2014; Goel and Singh 2015; Olascoaga and Riquelme 2017; Yusuf et al. 2018). Most studies illustrated a statistically significant interaction between the level of training and mini-CEX score and proved that with the increasing levels of competence, either as a result of greater length of training or progression over time, the scores of mini-CEX increased significantly. This is true for both overall clinical competence score and mean calculated average score of the specific domains. A few studies however failed to show such a relationship. Durning and colleagues, for example, stated that mean scores on the seven consecutive mini-CEXs were not significantly different (Durning et al. 2002). Some studies were conducted over a short span of time and hence did not show significant improvement in performance (Chang et al. 2013; Gupta et al. 2017).

On the other hand, some papers categorized trainees into two or more groups based on their performance in a criterion exam and then compared the mean scores of students in these groups. In these cases, studies showed that the ability of mini-CEX to discriminate is acceptable. Kogan and colleagues showed that students with "Honors" scored higher than those with "Pass" (Kogan et al. 2003). Hatala and collogues showed that the mean overall clinical competence score of students who successfully passed the RCPSC (Royal College of Physicians and Surgeons of Canada) Internal Medicine exam was significantly higher than those who failed (Hatala et al. 2006).

Correlation among competencies (construct validity)

The correlation among the components of the competence and the correlation among 6 competencies and the overall competence were extensively reported in the retrieved papers, showing high levels of correlation with statistical significance in both cases. Figures range between 0.51 (Berendonk et al. 2018) and 0.97 (Fernández Galvez 2011) for correlation coefficient among components and between 0.60 (Ergin et al. 2013) and 0.90 (Norcini et al. 1995) for correlation among competencies and the overall competence.

Factor analysis (construct validity)

Four studies performed a factor analysis and concluded that the results proved a single factor solution (Hill et al. 2009; Cook et al. 2010; Olascoaga and Riquelme 2017; Berendonk et al. 2018). One study indicated that the competency domains reflect a single pure latent variable, accounting for 66.5% of score variance (Hill et al. 2009). Another study looked at two data sets and noted that a single-factor solution existed which explained 100% of the variance (Cook et al. 2010). In another study, the single factor found in factor analysis accounted for 67% of the variance. All competency domains had a high loading, ranging from 0.58 to 0.84 (Olascoaga and Riquelme 2017). A recent study in Switzerland showed that one underlying factor explained 56% of the variance in supervisor scores. Factor loadings of the six domains on this single factor ranged from 0.58 to 0.84. A case sensitivity analysis was performed by excluding all mini-CEXs with one or more missing values. This reduced the sample and naturally the loadings, but still, there was a single factor solution (Berendonk et al. 2018).

Correlation between mini-CEX scores and other exams (criterion validity)

Assessment of the relationship between the mini-CEX scores and the results of well-established examinations, the result of which are often considered valid, has been reported in several papers and revealed promising findings. Durning et al. assessed the correlation between overall clinical competence mean scores and corresponding ABIM's Monthly Evaluation Form sections and overall Postgraduate Year two in-Training Examination percentile score. They found significant correlations with values as high as 0.57 (Durning et al. 2002). Kogan et al. studied the correlation of mini-CEX scores with exam scores, write-ups, inpatient and outpatient course grades, as well as final course grades. Correlations were significant in all cases and the greatest "*r*" was for inpatient course grades ($r = 0.47$) (Kogan et al. 2003). The RCPSC oral, bedside, and written scores were also significantly correlated with overall clinical competence score in another study (Hatala et al. 2006). OSCE, traditional clinical examination, Family/Internal Medicine Standardized Patient Exam, and Clinical Skills Exam were also used for this purpose in other studies, all finding significant correlations (Ney et al. 2009; Karanth et al. 2015; Rogausch et al. 2015; Humphrey-Murto et al. 2018). Humphrey-Murto et al. in their study on 3rd-year medical students concluded that the mini-CEX was not correlated with the written examination scores for any of the clerkship disciplines (Humphrey-Murto et al. 2018).

Qualitative data

Some studies have used qualitative methodology in the form of surveys and interviews to evaluate the perceptions and experiences of stakeholders. Hill and Kendall conducted semi-structured individual interviews and four group interviews with staff and students. Participants perceived the mini-CEX to be a more valid method than long case. However, they had concerns about the possibility of compensation between components (Hill and Kendall 2007). In a survey conducted by Weller, 55% of trainees and 58% of specialists believed the mini-CEX is a good measure of performance (Weller, Jones, et al. 2009). In a recent study by Khalil, however, 50% of residents and assessors were unsure whether mini-CEX was a valid method of assessment (Khalil et al. 2017).

Acceptability

Acceptability of the mini-CEX was assessed in 36 articles, mostly by assessing satisfaction of trainees and supervisors (Table 4).

Satisfaction via questionnaire

Most studies used an item at the end of the mini-CEX form to be rated by assessor and trainee immediately after the encounter while a few studies administered a separate questionnaire after the study period. In three studies, a 10-point scale was used (Weller, Jones, et al. 2009; Jackson and Wall 2010; Abadie et al. 2015), in three other studies a 5-point scale was administered (Singh and Sharma 2010; Weston and Smith 2014; Joshi et al. 2017), and in the rest of studies satisfaction was rated on a 9-point scale. Almost all studies also provided a comment box to collect open-ended responses. Except for two studies which reported medians (Weston and Smith 2014; Gupta et al. 2017), and one which illustrated the frequency distribution of the responses to each score (Brazil et al. 2012), all other studies reported the mean trainee and/or faculty satisfaction. Among 23 studies which reported faculty satisfaction, mean score ranged from 6.0 (Holmboe et al. 2004) to 9.0 (Olascoaga and Riquelme 2017) on a 9-point scale. With regard to trainees' satisfaction, only one study (Jackson and Wall 2010) found a mean resident satisfaction of 3.87 (on a 10-point Likert scale), which was largely different from all other studies. Excluding this outlier result, mean trainees' satisfaction ranged from 6.0 (Holmboe et al. 2004) to 8.8 (Alves de Lima et al. 2007) on a 9-point Likert scale. A study on using the mini-CEX for IMGs showed that about half of the examinees and almost all examiners were "satisfied" or "very satisfied" with the mini-CEX as a learning tool and were positive about the exam (Nair et al. 2008). When asked if the mini-CEX was a useful part of their training, foundation year 1 doctors in a study in the UK answered with a median score of 2.5 in a 5-point scale (Weston and Smith 2014). In one study, around 70 percent of trainees believed that the min-CEX was "an unrealistic reflection of their performance" (Jackson and Wall 2010). In yet another study, faculty raised the concern that mini-CEX may create bias as "it does not mimic real-life" (Gupta et al. 2017).

Satisfaction in the qualitative data

In general, results of the qualitative studies showed that mini-CEX was considered a very useful assessment instrument and a useful teaching tool (Alves de Lima et al. 2010), and overall satisfaction was high among both trainees and assessors (Brazil et al. 2012). They believed that it is a valuable assessment strategy and can be an adjunct to in-training assessment (Alves de Lima et al. 2005; Brazil et al. 2012; Gupta et al. 2017; Joshi et al. 2017), liked the realism (Alves de Lima et al. 2010), felt quite comfortable with it (Alves de Lima et al. 2005), believed they were judged fairly (Eriksen et al. 2009; Hill et al. 2009), and preferred it over long case as a summative exam (Hill and Kendall 2007). On the other hand, some studies have shed light on the drawbacks of mini-CEX. In one study, only 26% of trainees viewed the tool as a useful means of gaining feedback (Jackson and Wall 2010). In another study, the participant believed that there was a conflict between assessment and educational role of the min-CEX (Malhotra et al. 2008). Some also found mini-CEX to be anxiety provoking (Malhotra et al. 2008; Khalil et al. 2017). In a different study, though, only one quarter of the trainees believed that mini-CEX induced anxiety in them. In a study by Hill, some examiners found it hard to determine the necessary standard (Hill et al. 2009). In a study in Singapore, undergraduate medical trainees felt that their performance was compared unfairly against more senior students, but examiners felt otherwise (Yanting et al. 2016).

Educational impact

The educational impact of the mini-CEX has been evaluated in 30 studies (Table 4).

Assessing change in the performance of trainees

We found that only three studies actually investigated how the implementation of the mini-CEX could change the learners' clinical competence. One study used a modified Objective Structured Long Examination Record (OSLER) to assess internal medicine and neurology residents before and after the mini-CEX. The OSLER results of the mini-CEX group were significantly higher than baseline group in internal medicine, while no significant difference was found in neurology (Suhoyo et al. 2014). Another study used an experimental design to compare the students' competence as measured by the traditional clinical examination in the intervention and control groups. They found a statistically significant difference in the traditional clinical examination scores between the mini-CEX and control groups (52.93 vs. 47.72). Yet, this quite small difference might not be considered educationally significant. These investigators also compared the students' self-assessment scores before and after the intervention. The self-assessment scores also changed significantly (66.34 vs. 46.05) after implementation of the mini-CEX (Karanth et al. 2015). Kim and colleagues evaluated the impact of implementation of a mini-CEX requirement across all 3rd-year clerkships on clinical skills of the trainees by comparing the failure rate of students in an end-of-3rd-year 8-station summative OSCE. They noticed that the failure rate was significantly lower (three of 121 students [2%] vs. 14 of 114 students [12%], $p < 0.0046$) after the intervention (Kim et al. 2016).

Exploring trainees' and faculty's opinions

When asked about the value of the mini-CEX and its impact on learning, respondents were generally positive about the educational consequences of the mini-CEX. In one study, for example, participants found this tool as a valuable educational tool which helped them identify their strength and weaknesses (Hill and Kendall 2007). In a study by Alves de Lima, trainees believed that the mini-CEX promoted their deep learning approach and enhanced self-regulated learning (Alves de Lima et al. 2005). Some learners were in favor of the mini-CEX because they felt that it provided insight into their clinical competence and prepared them for successful completion of national exam (Malhotra et al. 2008). Faculty members also mentioned some beneficial impacts of the mini-CEX including promoting a more intensive interaction with residents (Alves de Lima et al. 2010) and enhancing the collegial relationship (Nair et al. 2008). When talking about the educational impact of the mini-CEX, feedback was frequently mentioned. Some believed that use of the mini-CEX created a strong culture of feedback (Nair et al. 2008) and others pointed out that it facilitated direct observation (Kim et al. 2016) as well as timely and specific feedback (Brazil et al. 2012). In one study, however, participants maintained that the mini-CEX tool tended not to motivate a change in practice (Jackson and Wall 2010). In a recent study in India, residents mentioned that the mini-CEX improved their clinical skills (72.4%), uplifted the personal development (62.0%), and imparted a better one to one student–teacher interaction (62.0%). Surprisingly, even 61.5% of faculty found that the mini-CEX was also useful for improved learning for themselves (Gupta et al. 2017). Yusuf et al. conducted a thematic analysis of the interviews and found that the reasons for improvement in scores were feedback, motivation, self-directed learning, and peer assisted learning (Yusuf et al. 2018)

The frequency and content of feedback

Eight studies tried to quantify the type and amount of delivered feedback through either audio-taping the feedback session or analysis of the written feedback comments on the returned assessment forms. Regarding the quantity and quality of feedback, the findings of the studies yielded conflicting results. Pernar showed that the use of the mini-CEX resulted in a twice as many qualitative feedback comments compared to the global assessment (Pernar et al. 2011). Many studies found that a kind of feedback had been provided in the majority of mini-CEX encounters (96% in Eriksen et al. 2009; 85.3% in Lin et al. 2012; 74.9% in Liao et al. 2013; and 92.6% in Chang et al. 2017), but the accuracy and usefulness of the feedback provided was questioned by many studies. One study, for instance, found that in 22.7% of cases the positive aspects of performance were not identified and in 28.2% of cases no suggestion for development was made; and still in 49.7% of cases no plan of action had been developed (Fernando et al. 2008). Action plan for improvement was mentioned to be present in 11% of the mini-CEXs in one study (Holmboe et al. 2004) and in 39.2% of the sessions in another study (Lin et al. 2012). A study by Chang et al. found that only 22.9% of all mini-CEXs contained all three components of feedback, namely, positive feedback, suggestions for development and agreed action plan (Chang et al. 2017). Liao and colleagues reported that among the 863 forms, 74.9% provided proper feedback (Liao et al. 2013). The skills that were the focus of feedback varied: in one study the domain of clinical judgment received most attention, with about half of the feedback items pertaining to this skill (Lin et al. 2012), and in another study clinical skills and medical knowledge were mainly (79.2 and 55.3%, respectively) addressed in feedbacks, but attitudes/professionalism were less frequently (22.7%) dealt with in feedbacks (Liao et al. 2013). A study in Switzerland looked for alignment of learning needs identified during the encounter with the learning goals mutually agreed on by the trainee and the supervisor at the end of the feedback session. They concluded that in spite of the fact that the supervisor and the trainee often identified similar learning needs, the majority of the mini-CEXs did not lead to *aligned* learning goals if at all (Montagne et al. 2014).

Cost and feasibility

Thirty-eight studies reported some data related to cost and feasibility (Table 4).

Cost

Only one of the papers, conducted in Australia, stated an annual cost in dollars simply by transforming the total annual faculty time expenditure into dollars. This study showed that total time of all the mini-CEXs over the study period was 36.51 hours in each of the 10 rotations in 1 year, representing an annual total of 365 hours, which would cost $A 80,000 per year. This was the only study that explicitly declared its primary aim as determining the cost and feasibility of adding the mini-CEX to current assessments (Brazil et al. 2012).

Time of the encounter

Half of the included studies ($N = 29$) reported either the total encounter time or the observation and feedback times separately. Some even reported separate data for different groups of assessors (e.g., residents vs. faculty). Time was reported as mean in some studies and as median in some others. Among 21 articles which reported mean observation time, the values ranged from 12.30 (Joshi et al. 2017) to 46.5 (Olascoaga and Riquelme 2017). The average time dedicated to feedback in the 19 articles which reported the value as mean varied from 5.73 (Kogan et al. 2003) to 20.1 (Weller, Jones, et al. 2009). Six articles only reported the total length of encounter, ranging from 17.7 (Norcini et al. 1997) to 31.50 (Norcini et al. 2003).

The completion rate

The proportion of completed mini-CEX forms to the preplanned required forms was obtained from 21 studies. The completion rate was generally high in most studies, ranging from 50.0% (Gupta et al. 2017) to 100% (Torre et al. 2007). Alves de Lima et al. in their large study on 17 cardiology residency programs in Argentina noticed that only 14.81% of the participants were evaluated four times or more during the 19-month study period. This number (four) was planned a priori as the minimum requirement (Alves de Lima et al. 2007). Fernández Galvez in his study on pediatrics residents in Argentina was able to obtain 22 assessment forms per student instead of the pre-planned four (Fernández Galvez 2011). In a 3-year study on undergraduate medical students in Australia, during the longitudinal integrated clerkship, many more mini-CEXs were submitted than required. The excess rate ranged from 5% to 180% (Playford et al. 2013). In a 2-year study on pediatric residents in Argentina, investigators defined feasibility as “the possibility of carrying out four assessments in at least 70% of the participants” and that “observations would be carried out in all the areas where the students rotate.” Analysis of the forms showed that 79.5% of the participants had at least 4 observations, made in all areas where students rotate (Urman et al. 2011).

Incomplete data

Chang et al. mentioned that in paper-based mini-CEX data gathering was incomplete for some dimensions. However, when the computer-based format was applied no missing data existed (Chang et al. 2013).

Perceptions and experiences of the stakeholders

In some studies, participants were asked if they expected difficulty arranging the mini-CEX. In one study on IMGs, 10 out of 16 trainees and 15 from 18 examiners stated that they never or only occasionally experienced difficulty (Nair et al. 2008). In another study, less than 30% of the participants had difficulty arranging mini-CEX (Weller, Jolly, et al. 2009). In the latter study, conducted on anesthesiology residents, less than 10% believed that the mini-CEX “slowed down the operating list” (Weller, Jolly, et al. 2009). In Yanting’s survey, participants believed that effective administration of mini-CEX was limited by “inter-tutor variability” and “lack of time” (Yanting et al. 2016). When asked through a questionnaire, some of the faculty members reported that the mini-CEX was “time-consuming.” Both residents and faculty members agreed that this form of assessment was “easy to carry out and sample wider areas” (Joshi et al. 2017). In qualitative evaluations, it was generally noted that students and teachers felt that the mini-CEX was feasible in most clinical settings (Goel and Singh 2015). In one study, however, faculty were skeptical of whether the mini-CEX would be possible for a large group of students (Gupta et al. 2017). Few participants considered the mini-CEX as an administrative burden in busy clinical workplaces (Castanelli et al. 2016). A study by Hill and colleagues revealed accounts of quality control issues. These included using other students or members of staff as “patients,” encounters with multiple patients, and failure to observe students (Hill et al. 2009).

Extra evidence

As we mentioned earlier, we also examined nine abstracts for which the full text did not exist or was not available. Three papers were on educational impact, four papers on feasibility, four papers on acceptability, and only one dealt with validity (in terms of content validity). Overall, the quality of the studies was not high, and the information obtained did not add to the findings of the full-text articles. For instance, the observation and feedback time stated in these papers were within the range that the already included studies reported.

Discussion

The aim of this study was to systematically identify and synthesize the evidence relating to the utility of the mini-CEX in undergraduate and postgraduate medical programs. In order to investigate all aspects of utility, we used the framework proposed by van der Vleuten consisting of validity, reliability, educational impact, acceptability, and the cost of the assessment tool.

Main findings

What is the reliability of the mini-CEX in the assessment of undergraduate and postgraduate medical trainees?

The findings of this review revealed different values for reliability coefficient. Cronbach’s alpha has consistently been reported to be high and meta-analysis of the pooled data corroborated this finding. It is worth noting that the number of encounters in the included studies fluctuated across the studies and the estimated reliability or generalizability coefficients should, therefore, be considered in light of this moderator. In addition, whether the overall mini-CEX score or the composite score (the average of domain scores) was used for calculation also impacts the interpretation of the findings. Furthermore, most studies which conducted G analysis failed to describe the precise design for the analysis. Whether assessors are nested or crossed with students affects the generalizability coefficient. Hypothetically, several possibilities can be imagined: a number of observers rating one single physician-patient encounter, several physician patient encounters each rated by one single observer, and any combination thereof. In reality, however, each encounter is often observed by one evaluator.

We found out that the contributions of facets other than the trainees' competency (such as assessor stringency, trainee ability, and assessor trainee-related subjectivity) in score variance were concerning. The number of encounters required to achieve a reliability coefficient of 0.7 varied in different studies, with numbers as small as five in one study and 60 in another investigation. While the commonly looked-for G coefficient may require a relatively high number of observations, Norcini argues that "CI provides additional information that permits test length to be tailored to specific situations" (Norcini et al. 1995). In other words, even smaller number of encounters will result in SEMs and CIs of enough precision for most assessment purposes. In a 9-point Likert scale it is wise to accept a CI of ±0.5 or a SEM of ±0.26. Considering these benchmarks, a smaller number of mini-CEXs (in the range of 4–10) could be considered precise yet feasible, even when it corresponds to a G coefficient of less than 0.7. Moreover, in the interpretation of SEM and G coefficient, the practical significance matters. While a wider CI would be adequate for a trainee who is far enough from the acceptable threshold, extra encounters should be arranged for trainees who are within borderline zones to achieve narrower SEMs and hence more precise decisions.

What is the validity evidence for the mini-CEX in the assessment of undergraduate and postgraduate medical trainees?

Since the mini-CEX was successfully administered in various settings and across a representative range of patient problems, it is reasonable to consider an acceptable content validity for the tool. None of the studies, however, referred to a pre-determined table of specifications and they all failed to objectively compare the list of clinical problem against the blueprint. Furthermore, the included studies provided evidence to support the construct validity of mini-CEX. Although not unanimously, the ability of the mini-CEX to discriminate between various levels of competency was established in the majority of studies. Overall and average domain scores were shown to increase, pointing to significant progress in performance as the trainees move along the course of training. This holds true when the scores of high-achievers are compared with those of less-competent learners. Ansari et al. in a systematic review of the studies on the validity of mini-CEX identified 11 papers and conducted meta-analysis on their data. One group of studies in their review investigated the construct validity of mini-CEX and showed that a change in the year of residency training was manifested in the difference in performance in the mini-CEX. They concluded that the mini-CEX has evidence of construct validity when used with residents across the years of a residency program. Residents' performance on the mini-CEX items across 1 year of residency training showed "small" to "medium" effect size differences. They calculated a combined fixed-effect and random-effects size for the overall clinical competence item of "medium" magnitude [$d = 0.50$; 95% CI, 0.31–0.70] (Ansari et al. 2013).

Another group of studies still looking for construct validity evidence showed that the differences between performance level within a peer group led to a mean difference in clinical performance on three items and a total mean score of the mini-CEX. The effect size in these studies varied but was generally high. The effect size differences between performance levels within a peer group (superior/honors, marginal/high pass, poor/pass) ranged from $d = 0.43$ (95% CI, 0.23–0.63) in one study on the total mean score of the mini-CEX up to $d = 1.86$ (95% CI, 0.31–3.40) on the physical examination skills item (Ansari et al. 2013).

All four factor analytic studies which explored score dimensionality revealed that a single latent dimension, namely global clinical performance, accounts for the majority, if not all of the score variance. This might be attributed to rating biases such as halo effect, breach in the underlying domain theory, or a technical flaw in the assessment tool. Since this finding suggests that specific competency domains contribute little to the score variance and may not offer a valid discrimination among unique aspects of clinical abilities, the domain scores should be interpreted cautiously.

Evidence for the criterion validity of the mini-CEX is supported by the moderate to high magnitude of correlations between the mini-CEX scores and other clinical skill achievements. In the review by Ansari et al. five studies looked at the predictive/concurrent validity of the mini-CEX by comparing the trainees' mini-CEX ratings with some other criterion measure. They found that the combined random-effects size for the overall clinical competence item was "medium" [$d = 0.64$; 95% CI, 0.48–0.77]. They concluded that the mini-CEX shows evidence of criterion-related validity when compared with other clinical skill achievement (e.g., certifying oral and written examinations) or performance (e.g., in-training evaluation reports, inpatient or outpatient write-ups) measures. They found "small" to "large" correlation coefficients with combined effect sizes (Ansari et al. 2013).

Hawkins and colleagues conducted a literature review in which they applied Kane's framework to synthesize the evidence. They found evidence for extrapolation component of the validity argument by showing that mini-CEX outcomes are related to the construct of interest, as measured by other clinical skills assessment tools. They also argued that the weakest component of the mini-CEX validity argument seems to be in the area of scoring. In terms of the scoring component, three issues are of primary concern: high inter-item correlations, rater selection and training, and leniency (Hawkins et al. 2010).

How acceptable is the mini-CEX to medical students and faculty members in undergraduate and postgraduate settings?

Acceptability is viewed as a core concern of any assessment tool. Successful implementation requires that all stakeholders get along with the process. Overall, except for outlier results of one study, the findings of this review revealed evidence in favor of high satisfaction rates among students and assessors. In the qualitative studies focusing on the acceptability of the mini-CEX comments were generally positive. However, a small number of participants raised concerns regarding realism and fairness of this assessment tool, as well as the amount of anxiety it provoked. On the other hand, we found studies in which participants liked the realism and considered the mini-CEX as a fair method of assessment. Some also mentioned that a conflict might exist between the assessment and the educational role of the encounter.

The high level of satisfaction with the mini-CEX is not unexpected since it resembles the ordinary interaction between a trainee and a supervisor in the clinical setting. Furthermore, most studies have mentioned an orientation program for familiarizing the trainees and faculty with the purpose and process of evaluation, which might have positively affected the perception of the involved parties. However, when interpreting this evidence, the tendency of the respondents to answer questions in a way that will please the investigators, known as response bias, should be considered.

Factors affecting satisfaction such as trainees' background, case difficulty, and the time of encounter have been the subject of some studies but were beyond the scope of this review. So is a comparison between acceptability of the mini-CEX and other assessment tools. While patients are an integral part of workplace-based assessments, all studies have relied on the trainees' and evaluators' opinions, and none have sought the patients' perspectives and addressed the question of how acceptable the mini-CEX is to the patients.

What is the educational impact of the mini-CEX on the undergraduate and postgraduate medical trainees?

The effect of the mini-CEX on students' learning has been demonstrated directly through observing the change in trainees' performance, or indirectly, through analysis of the feedback provided, or assessing stakeholders' perception of the mini-CEX as an educational tool. The findings of the three included studies that directly tackled the question of educational value revealed that implementation of the mini-CEX improved students' scores and decreased failure rates in other examinations such as OSCE, OSLER, or traditional clinical examination.

Feedback is an integral part of the mini-CEX and so several studies have sought for the evidence that ensures delivery of effective feedback. These studies have analyzed the frequency of feedback provision as well as the content of the feedback. The findings, though, were not invariably promising. Although feedback had been provided in the majority of the mini-CEX encounters in several studies, its quality and effectiveness were questionable. It was frequently observed that suggestions for development and agreed action plans were lacking in the feedback.

Lörwald and colleagues systematically reviewed 20 papers on the educational impact of mini-CEX and noticed that the majority of the articles had investigated effects of the mini-CEX on learners' reactions (Kirkpatrick level 1), showing mixed results (Lörwald, Lahner, Nouns, et al. 2018). In our study, we used the Van der Vleuten formula instead of the Kirkpatrick model and hence did not classify the learners' satisfaction as educational impact and considered it instead as evidence of acceptability. However, many studies asked either through conducting surveys or gathering qualitative data the trainees' and faculty's opinion regarding the educational impact of this assessment tool. We did include these data as evidence for educational impact, but the findings were not suitable for meta-analysis or qualitative synthesis. Therefore, we cannot offer any solid conclusion based on these studies and simply reported their findings. In Lörwald's study, three studies which had reported the effects of the mini-CEX on trainee performance (Kirkpatrick level 2b) underwent meta-analysis. A positive effect of the mini-CEX on trainee performance was revealed (Lörwald, Lahner, Nouns, et al. 2018). Lörwald and colleagues, like us, failed to find any evidence for change in attitudes or perceptions (Kirkpatrick level 2b), behavior (Kirkpatrick level 3), organizational practice (Kirkpatrick level 4a), or benefits to patients (Kirkpatrick level 4b). In a qualitative synthesis of systematically reviewed papers, these investigators analyzed the influencing factors on the educational impact of the mini-CEX. They found out that time for the mini-CEX, usability of the tools, supervisors' knowledge about how to use the mini-CEX, supervisors' attitude to the mini-CEX, trainees' knowledge about the mini-CEX, trainees' perception of the mini-CEX, observation, feedback, and trainees' appraisal of feedback can affect the educational impact of the mini-CEX (Lörwald, Lahner, Greif, et al. 2018). This was not among the objectives of our study and we did not deal with this issue in our review.

The reason for these inconsistent findings may be the that the outcome measures considered for the educational effect of the min-CEX were general and often took the form of the regular assessments at the end of a rotation or course. This could have resulted in the invisibility of a large proportion of the mini-CEX effect. Furthermore, the importance of the proper implementation of the mini-CEX, which has been emphasized in the literature (Durlak and DuPre 2008; Hauer et al. 2011), might have affected these findings. We noted that the evidence for this utility criteria is limited and is mainly focused on stakeholders' reactions and opinions, with only few included studies (none of them postgraduate training) in adopting a method to directly measure the educational impact of the mini-CEX. This can be an area for further study. While the current mini-CEX form provides a structured format for performing direct observation, it does not facilitate interactive feedback. A revised format may be required. This might also be an area for further investigation.

What is the cost and feasibility of using the mini-CEX in the assessment of undergraduate and postgraduate medical trainees?

We found only one study within the medical education literature specifically addressing the issue of cost. Even this single paper reported a monetary equivalent for faculty time dedicated to performing the mini-CEX. Other direct or indirect costs were addressed. The cost was reported globally and not as per assessee. No attempts have been made to evaluate the cost-effectiveness of this tool.

There were only a few papers explicitly reporting feasibility as an outcome. Therefore, we generated some indicators from what had been reported by authors. Time of the encounter, the completion rate (proportion of completed mini-CEX forms to the pre-planned required forms), and the rate of incomplete data were adapted as surrogate markers of feasibility. We also considered the survey questions and comments that directly asked about the practicability of the mini-CEX as feasibility evidence. We suggest that our feasibility framework can be applied for other assessment tools such as DOPS.

Based on the findings of the included studies in this review, there is ample evidence of the mini-CEX being feasible, as it has been implemented and evaluated in different settings. The encounter time in different studies varied, up to one hour in some cases. However, most studies reported durations similar to what the developers originally intended, which was about 20 min. The completion rates (when provided) were reported with mixed results, although generally above 50%. Qualitative studies and surveys revealed that only a small percentage of the participants had difficulty arranging mini-CEX and that only a minority of the stakeholders believed that the mini-CEX interfered with the routine process of care.

The strengths and limitations of the review

The main strength of this study is that we identified and recruited all evidence that evaluated the mini-CEX in terms of any psychometric or educational aspects and then synthesized the findings in a comprehensive model while other reviews on this assessment tool focused only on one particular feature.

Moreover, this systematic review benefits from a comprehensive and effective search syntax. In addition to conducting the search in seven electronic databases, we included gray literature and managed to contain non-English papers. Pervious reviews suffered from methodological flaws due to incomplete search strategy.

Heterogeneity in the contexts, data gathering, and data analysis of the included studies led to our inability to conduct a meta-analysis (with the exception of Cronbach’s alpha), which might affect the robustness of the educational recommendations driven from the findings of this study. The findings of this review should be interpreted regarding limitations of the primary studies. Some studies modified the original mini-CEX form in terms of its rating scale and domains. Some studies used overall scores, while others used domain scores for their analysis. Several studies either did not explicitly attribute the measured variables to one of the utility criteria or reported the same measures as evidence for different criteria. Hence, we have made an attempt to organize all of the variables and measures in a meaningful and consistent approach.

Implications for educators and researchers

From a practical standpoint, administrators and faculty members who would like to observe and evaluate medical trainees within workplace in a meaningful and credible manner can opt for the mini-CEX as an assessment tool with positive educational consequences. However, like other assessment tools, the mini-CEX as a single measure of competence suffers from some limitations and should be used in conjunction with other instruments. It should also be noted that none of the utility characteristics reported in different studies are inherent attributes of this assessment tool. Particular attention should be paid to proper implementation of this tool and establishment of a quality assurance system.

Having conducted a thorough analysis of the psychometric properties of the mini-CEX, we identified several gaps of knowledge in this field suggesting areas for further research. While several studies have delved into the effect of potential moderators on students’ scores, future investigations should focus on the effect of these moderators on the utility characteristics. This includes variables related to the rater (e.g., their experience and background), the mini-CEX form (e.g., competency domains and length of the scale), and the implementation process (e.g., using digital tool). The commonly reported issues with score inflation and grade restriction, as well as measures to overcome these problems are also worth studying. If the mini-CEX is to be incorporated more extensively into the assessment programs, its direct and indirect costs need further scrutiny. Future research may also be directed at investigating the effect of the mini-CEX on the changes in the behavior, organization of practice, and benefits to patients.

To enhance our understanding of the results of primary studies and to facilitate future reviews, rigorous reporting is essential. We recommend development of a framework above and beyond the usual reporting standards.

Conclusion

In summary, the mini-CEX is widely used as a formative and summative assessment tool and appears to have reasonable validity and reliability. The reported acceptability and feasibility should be interpreted in the light of the required number of encounters needed to achieve desired reliability. By providing a framework for structured observation and feedback, the mini-CEX bears a favorable educational impact.

Acknowledgments

The authors would like to thank the Alabama BICC (BEME International Collaborating Center) and the BREC (BEME Review Editorial Committee) for their comments and feedback on the study protocol and manuscript.

Disclosure statement

The authors declare that they have no competing interests. All authors confirmed the final version and given approval for the manuscript to be published. SMH, MJ, SN, MS, and RM are members of BEME International Collaboration Center at Tehran University of Medical Sciences (TUMS BICC).

Funding

This study has been funded and supported by Tehran University of Medical Sciences.

Notes on contributors

Sara Mortaz Hejri, MD, PhD, AFAMEE, Assistant Professor, Department of Medical Education, Health Professions Education Research Center, Tehran University of Medical Sciences, Tehran, Iran.

Mohammad Jalili, MD, Professor, Department of Emergency Medicine, Department of Medical Education, Tehran, Iran.

Rasoul Masoomi, PhD candidate in Medical Education, Department of Medical Education, Tehran University of Medical Sciences, Tehran, Iran.

Mandana Shirazi, PhD, Associate Professor, Education Development Center, Department of Medical Education, Tehran University of Medical Sciences, Tehran, Iran; Affiliated Associate Professor of Department of Clinical Science and Education at SOS Hospital, Karolina Institute, Stockholm, Sweden.

Saharnaz Nedjat, MD, PhD, Professor, Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran.

John Norcini, PhD, President Emeritus of Foundation for Advancement of International Medical Education and Research (FAIMER), Philadelphia, Pennsylvania, USA.

ORCID

Sara Mortaz Hejri http://orcid.org/0000-0002-9979-0529
Mohammad Jalili http://orcid.org/0000-0002-0689-5437
Rasoul Masoomi http://orcid.org/0000-0003-4818-6062
John Norcini http://orcid.org/0000-0002-8464-4115

References

Abadie Y, Battolla J, Zubieta A, Dartiguelongue J, Pascual C, Elías Costa C, Vassallo JC, Rodríguez S. 2015. Uso de descriptores durante la implementación de mini-cex en la residencia de pediatría. Medicina. 75(5):289–296.

Alves de Lima A, Barrero C, Baratta S, Castillo Costa Y, Bortman G, Carabajales J, Conde D, Galli A, Degrange G, Van DER Vleuten C. 2007. Validity, reliability, feasibility and satisfaction of the Mini-Clinical Evaluation Exercise (Mini-CEX) for cardiology residency training. Med Teach. 29(8):785–790.

Alves de Lima A, Conde D, Aldunate L, van der Vleuten C. 2010. Teachers' experiences of the role and function of the mini clinical evaluation exercise in postgraduate training. Int J Med Educ. 1: 68–73.

Alves de Lima A, Conde D, Costabel J, Corso J, Van der Vleuten C. 2013. A laboratory study on the reliability estimations of the mini-CEX. Adv in Health Sci Educ. 18(1):5–13.

Alves de Lima A, Henquin R, Thierer J, Paulin J, Lamari S, Belcastro F, Van der Vleuten CP. 2005. A qualitative study of the impact on learning of the mini clinical evaluation exercise in postgraduate training. Med Teach. 27(1):46–52.

Ansari A, Ali SK, Donnon T. 2013. The construct and criterion validity of the mini-CEX: a meta-analysis of the published research. Acad Med. 88(3):413–420.

Berendonk C, Rogausch A, Gemperli A, Himmel W. 2018. Variability and dimensionality of students' and supervisors' mini-CEX scores in undergraduate medical clerkships a multilevel factor analysis. BMC Med Educ. 18(1):100.

Brazil V, Ratcliffe L, Zhang J, Davin L. 2012. Mini-CEX as a workplace-based assessment tool for interns in an emergency department–-does cost outweigh value? Med Teach. 34(12):1017–1023.

Buckley S, Coleman J, Davison I, Khan KS, Zamora J, Malick S, Morley D, Pollard D, Ashcroft T, Popovic C, et al. 2009. The educational effects of portfolios on undergraduate student learning: a Best Evidence Medical Education (BEME) systematic review. BEME guide no. 11. Med Teach. 31(4):282–298.

Castanelli DJ, Jowsey T, Chen Y, Weller JM. 2016. Perceptions of purpose, value, and process of the mini-Clinical Evaluation Exercise in anesthesia training. Can J Anesth/J Can Anesth. 63(12):1345–1356.

Chang Y-C, Chen C-K, Chen J-C, Liao C-H, Lee C-H, Chen Y-C, Ng C-J, Huang J-L, Lee S-T. 2013. Implementation of the mini-clinical evaluation exercise in postgraduate Year 1 residency training in emergency medicine: Clinical experience at Chang Gung Memorial Hospital. J Acute Med. 3(3):110–115.

Chang YC, Lee CH, Chen CK, Liao CH, Ng CJ, Chen JC, Chaou CH. 2017. 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 in Health Sci Educ. 22(1):57–67.

Cook DA, Beckman TJ. 2009. Does scale length matter? A comparison of nine- versus five-point rating scales for the mini-CEX. Adv in Health Sci Educ. 14(5):655–664.

Cook DA, Beckman TJ, Mandrekar JN, Pankratz VS. 2010. Internal structure of mini-CEX scores for internal medicine residents: factor analysis and generalizability. Adv in Health Sci Educ. 15(5):633–645.

Cook DA, Bordage G, Schmidt HG. 2008. Description, justification and clarification: a framework for classifying the purposes of research in medical education. Med Educ. 42(2):128–133.

Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. 2009. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. J Gen Intern Med. 24(1): 74–79.

Cruess R, McIlroy JH, Cruess S, Ginsburg S, Steinert Y. 2006. The Professionalism Mini-evaluation Exercise: a preliminary investigation. Acad Med. 81(Suppl):S74–S78.

Donato AA, Pangaro L, Smith C, Rencic J, Diaz Y, Mensinger J, Holmboe E. 2008. Evaluation of a novel assessment form for observing medical residents: a randomised, controlled trial. Med Educ. 42(12):1234–1242.

Durlak JA, DuPre EP. 2008. Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. Am J Commun Psychol. 41(3–4):327–350.

Durning SJ, Cation LJ, Markert RJ, Pangaro LN. 2002. Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. Acad Med. 77(9):900–904.

Ergin S, Özdemir S, Büke AS, Cüneyt OK, Kaçar N. 2013. Mezuniyet öncesi dermatovenereloji eğitiminde mini klinik değerlendirme uygulaması: Pamukkale Üniversitesi Tıp Fakültesi'nin deneyimi. Arch Turkish Dermatol Venerol. 47(1):54–58.

Eriksen JG, Simonsen D, Bastholt L, Aspegren K, Vinther C, Kruse K, Kodal T. 2009. Mini clinical evaluation exercise as evaluation tool of communicative and cooperative skills in the outpatient clinic. Ugeskr Laeger. 171(12):1003–1006.

Ferenchick GS, Foreback J, Towfiq B, Kavanaugh K, Solomon D, Mohmand A. 2010. The implementation of a mobile problem-specific electronic CEX for assessing directly observed student-patient encounters. Med Educ Online. 29:15.

Fernández Galvez GM. 2011. Evaluación de las competencias clínicas en una residencia de pediatría con el Mini-CEX (Mini-Clinical Evaluation Exercise). Arch Argentinos de Pediatria. 109(4):314–320.

Fernando N, Cleland J, McKenzie H, Cassar K. 2008. Identifying the factors that determine feedback given to undergraduate medical students following formative mini-CEX assessments. Med Educ. 42: 89–95.

Gade SA, Chari SN, Chalak A. 2017. Use of mini-CEX as a teaching learning method in physiology for undergraduate medical students. Natl J Physiol Pharm Pharm. 7(4):1–5.

Gingerich A, Ramlo SE, van der Vleuten CPM, Eva KW, Regehr G. 2017. Inter-rater variability as mutual disagreement: identifying raters' divergent points of view. Adv in Health Sci Educ. 22(4):819–838.

Goel A, Singh T. 2015. The usefulness of Mini Clinical Evaluation Exercise as a learning tool in different pediatric clinical settings. Int J Appl Basic Med Res. 5(4):32–34.

Gohel M, Singh US, Bhanderi D, Phatak A. 2016. Developing and pilot testing of a tool for "clinicosocial case study" assessment of community medicine residents. Educ Health. 29(2):68–74.

Golnik KC, Goldenhar L, Gittinger JW, Jr, Lustbader JM. 2004. The Ophthalmic Clinical Evaluation Exercise (OCEX). Ophthalmology. 111(7):1271–1274.

Gordon M, Gibbs T. 2014. STORIES statement: publication standards for healthcare education evidence synthesis. BMC Med. 12(1):143.

Gupta S, Sharma M, Singh T. 2017. The acceptability and feasibility of mini-clinical evaluation exercise as a learning tool for pediatric postgraduate students. Int J App Basic Med Res. 7(5):19.

Han PK, Keranen LB, Lescisin DA, Arnold RM. 2005. The palliative care clinical evaluation exercise (CEX): an experience-based intervention for teaching end-of-life communication skills. Acad Med. 80(7): 669–676.

Hatala R, Ainslie M, Kassen BO, Mackie I, Roberts JM. 2006. Assessing the mini-clinical evaluation exercise in comparison to a national specialty examination. Med Educ. 40(10):950–956.

Hauer KE, Holmboe ES, Kogan JR. 2011. Twelve tips for implementing tools for direct observation of medical trainees' clinical skills during patient encounters. Med Teach. 33(1):27–33.

Hawkins RE, Margolis MJ, Durning SJ, Norcini JJ. 2010. Constructing a validity argument for the mini-clinical evaluation exercise: a review of the research. Acad Med. 85(9):1453–1461.

Higgins, JPT, Green, S., editors. 2011. Cochrane handbook for systematic reviews of interventions. Version 5.1.0 [updated March 2011]. The Cochrane Collaboration. Available from http://handbook.cochrane.org.

Hill F, Kendall K. 2007. Adopting and adapting the mini-CEX as an undergraduate assessment and learning tool. Clin Teach. 4(4): 244–248.

Hill F, Kendall K, Galbraith K, Crossley J. 2009. Implementing the undergraduate mini-CEX: a tailored approach at Southampton University. Med Educ. 43(4):326–334.

Holmboe ES, Huot S, Chung J, Norcini J, Hawkins RE. 2003. Construct validity of the miniclinical evaluation exercise (miniCEX). Acad Med. 78(8):826–830.

Holmboe ES, Yepes M, Williams F, Huot SJ. 2004. Feedback and the mini clinical evaluation exercise. J Gen Intern Med. 19(5 Pt 2): 558–561.

Humphrey-Murto S, Côté M, Pugh D, Wood TJ. 2018. Assessing the validity of multidisciplinary mini-clinical evaluation exercise. Teach Learn Med. 30(2):152–161.

Jackson D, Wall D. 2010. An evaluation of the use of the mini-CEX in the foundation programme. Br J Hosp Med. 71(10):584–588.

Joshi MK, Singh T, Badyal DK. 2017. Acceptability and feasibility of mini-clinical evaluation exercise as a formative assessment tool for workplace-based assessment for surgical postgraduate students. J Postgrad Med. 63(2):100–105.

Karanth KL, Kanagasabai S, Ibrahim SB, Najimuddin M, Marasinghe DK, De S. 2015. Structured program for final-year undergraduate students to improve clinical skills to prepare for effective patient management. Internet J Gynecol Obstet. 19:1–9.

Khalil S, Aggarwal A, Mishra D. 2017. Implementation of a mini-clinical evaluation exercise (mini-CEX) program to assess the clinical competence of postgraduate trainees in pediatrics. Indian Pediatr. 54(4): 284–287.

Kim S, Willett LR, Noveck H, Patel MS, Walker JA, Terregino CA. 2016. Implementation of a mini-CEX requirement across all third-year clerkships. Teach Learn Med. 28(4):424–431.

Kogan JR, Bellini LM, Shea JA. 2003. Feasibility, reliability, and validity of the mini-clinical evaluation exercise (mCEX) in a medicine core clerkship. Acad Med. 78(Supplement):S33–S5.

Kogan JR, Holmboe ES, Hauer KE. 2009. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. JAMA. 302(12):1316–1326.

Liao KC, Pu SJ, Liu MS, Yang CW, Kuo HP. 2013. Development and implementation of a mini-clinical evaluation exercise (mini-CEX) program to assess the clinical competencies of internal medicine residents: from faculty development to curriculum evaluation. BMC Med Educ. 26(13):31.

Lie D, Encinas J, Frances S, Michael P. 2010. Do faculty show the 'halo effect' in rating students compared with standardized patients during a clinical examination? Internet J Fam Practice. 8(2):1–8.

Lin C, Chiu T, Yen D, Chong C. 2012. Mini-clinical evaluation exercise and feedback on postgraduate trainees in the emergency department: a qualitative content analysis. J Acute Med. 2(1):1–7.

Lörwald AC, Lahner FM, Greif R, Berendonk C, Norcini J, Huwendiek S. 2018. Factors influencing the educational impact of Mini-CEX and DOPS: a qualitative synthesis. Med Teach. 40(4):414–420.

Lörwald AC, Lahner FM, Nouns ZM, Berendonk C, Norcini J, Greif R, Huwendiek S. 2018. The educational impact of mini-clinical evaluation exercise (mini-CEX) and direct observation of procedural skills (DOPS) and its association with implementation: a systematic review and meta-analysis. PLoS One. 13(6):e0198009.

Malhotra S, Hatala R, Courneya CA. 2008. Internal medicine residents' perceptions of the mini-clinical evaluation exercise. Med Teach. 30(4):414–419.

Margolis MJ, Clauser BE, Cuddy MM, Ciccone A, Mee J, Harik P, Hawkins RE. 2006. Use of the mini-clinical evaluation exercise to rate examinee performance on a multiple-station clinical skills examination: a validity study. Acad Med. 81(Suppl):S56–S60.

Meresh E, Daniels D, Sharma A, Rao M, Mehta K, Schilling D. 2018. Review of mini-clinical evaluation exercise (mini-CEX) in a psychiatry clerkship. AMEP. 9:279–283.

Miller A, Archer J. 2010. Impact of workplace based assessment on doctors' education and performance: a systematic review. BMJ. 341:c5064.

Mills E, Blenkinsopp A, McKinley RK, Black P. 2011. The assessment of observed practice: a literature review. Keele: University of Keele.

Montagne S, Rogausch A, Gemperli A, Berendonk C, Jucker-Kupper P, Beyeler C. 2014. The mini-clinical evaluation exercise during medical clerkships: are learning needs and learning goals aligned? Med Educ. 48(10):1008–1019.

Mortaz Hejri S, Jalili M, Shirazi M, Masoomi R, Nedjat S, Norcini J. 2017. The utility of mini-Clinical Evaluation Exercise (mini-CEX) in undergraduate and postgraduate medical education: protocol for a systematic review. Syst Rev. 6(1):146.

Musick DW, Bockenek WL, Massagli TL, Miknevich MA, Poduri KR, Sliwa JA, Steiner M. 2010. Reliability of the physical medicine and rehabilitation resident observation and competency assessment tool: a multi-institution study. Am J Phys Med Rehab. 89(3): 235–244.

Nair BR, Alexander HG, McGrath BP, et al. 2008. The mini clinical evaluation exercise (mini-CEX) for assessing clinical performance of international medical graduates. Med J Aust. 189(3):159–161.

Ney EM, Shea JA, Kogan JR. 2009. Predictive validity of the mini-clinical evaluation exercise (mCEX): do medical students' mCEX ratings correlate with future clinical exam performance? Acad Med. 84(Supplement):S17–S20.

Norcini JJ, Blank LL, Arnold GK, Kimball HR. 1995. The mini-CEX (clinical evaluation exercise): a preliminary investigation. Annals of Internal Medicine. 123(10):795–799.

Norcini JJ, Blank LL, Arnold GK, Kimball HR. 1997. Examiner differences in the mini-CEX. Adv Health Sci Educ Theory Practice. 2(1):27–33.

Norcini JJ, Blank LL, Duffy FD, Fortna GS. 2003. The mini-CEX: a method for assessing clinical skills. Ann Intern Med. 138(6):476–481.

Olascoaga AC, Riquelme A. 2017. Aplicación longitudinal del Mini Clinical Examination (Mini-CEX) enmédicos residentes. Educ Med. 20:25–28.

Paravicini I, Peterson CK. 2015. Introduction, development, and evaluation of the miniclinical evaluation exercise in postgraduate education of chiropractors. J Chiropractic Educ. 29(1):22–28.

Pelgrim EAM, Kramer AWM, Mokkink HGA, van den Elsen L, Grol RP, van der Vleuten C. 2011. In-training assessment using direct observation of single-patient encounters: a literature review. Adv Health Sci Educ. 16(1):131–142.

Pernar LIM, Peyre SE, Warren LEG, Gu X, Lipsitz S, Alexander EK, Ashley SW, Breen EM. 2011. Mini-clinical evaluation exercise as a student assessment tool in a surgery clerkship: lessons learned from a 5-year experience. Surgery. 150(2):272–277.

Playford D, Kirke A, Maley M, Worthington R. 2013. Longitudinal assessment in an undergraduate longitudinal integrated clerkship: the mini Clinical Evaluation Exercise (mCEX) profile. Med Teach. 35(8):e1416–e1421.

Pottier P, Cohen AF, Steichen O, Desprets M, Pha M, Espitia A, Georgin-Lavialle S, Morel A, Hardouin JB. 2018. Validité et reproductibilité de deux grilles d'observation des compétences cliniques des internes en DES de médecine interne. La Rev Méd Interne. 39(1):4–9.

Rodriguez MC, Maeda Y. 2006. Meta-analysis of coefficient alpha. Psychol Methods. 11(3):306–322.

Rogausch A, Beyeler C, Montagne S, Jucker-Kupper P, Berendonk C, Huwendiek S, Gemperli A, Himmel W. 2015. The influence of students' prior clinical skills and context characteristics on mini-CEX scores in clerkships–a multilevel analysis. BMC Med Educ. 15(1):208.

Saeed N, Tariq N, Jaffery T. 2015. Mini-CEX (clinical evaluation exercise) as an assessment tool at Shifa College of Medicine, Islamabad, Pakistan. Rawal Med J. 40(2):220–224.

Sandilands DD, Zumbo BD. 2014. (Mis)Alignment of medical education validation research with contemporary validity theory: the mini-CEX as an example. In: Zumbo BD, Chan EKH, editors. Validity and validation in social, behavioral, and health sciences. Cham: Springer. p. 289–310.

Singh T, Sharma M. 2010. Mini-clinical examination (CEX) as a tool for formative assessment. The National Medical Journal of India. 23(2): 100–102.

Suhoyo Y, Schönrock-Adema J, Rahayu GR, Kuks JB, Cohen-Schotanus J. 2014. Meeting international standards: a cultural approach in implementing the mini-CEX effectively in Indonesian clerkships. Med Teach. 36(10):894–902.

Torre DM, Simpson DE, Elnicki DM, Sebastian JL, Holmboe ES. 2007. Feasibility, reliability and user satisfaction with a PDA-based mini-CEX to evaluate the clinical skills of third-year medical students. Teach Learn Med. 19(3):271–277.

Torre DM, Treat R, Durning S, Elnicki DM. 2011. Comparing PDA- and paper-based evaluation of the clinical skills of third-year students. Wilkinson Med J. 110(1):9–13.

Urman G, Folgueral S, Gasparri M, et al. 2011. Assessment of competence in pediatric postgraduate education: implementation of a pediatric version of the mini-CEX. Arch Argent Pediat. 109(06): 492–498.

Van der Vleuten C. 1996. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ. 1(1):41–67.

Weller JM, Castanelli DJ, Chen Y, Jolly B. 2017. Making robust assessments of specialist trainees' workplace performance. Br J Anaesth. 118(2):207–214.

Weller JM, Jolly B, Misur MP, Merry AF, Jones A, Crossley JGM, Pedersen K, Smith K. 2009. Mini-clinical evaluation exercise in anaesthesia training. Br J Anaesth. 102(5):633–641.

Weller JM, Jones A, Merry AF, Jolly B, Saunders D. 2009. Investigation of trainee and specialist reactions to the mini-clinical evaluation exercise in anaesthesia: implications for implementation. Br J Anaesth. 103(4):524–530.

Weller JM, Misur M, Nicolson S, Morris J, Ure S, Crossley J, Jolly B. 2014. Can I leave the theatre? A key to more reliable workplace-based assessment. Br J Anaesth. 112(6):1083–1091.

Weston PS, Smith CA. 2014. The use of mini-CEX in UK foundation training six years following its introduction: lessons still to be learned and the benefit of formal teaching regarding its utility. Med Teach. 36(2):155–163.

Yanting SL, Sinnathamby A, Wang D, Heng MTM, Hao JLW, Lee SS, Yeo SP, Samarasekera DD. 2016. Conceptualizing workplace based assessment in Singapore: undergraduate mini-clinical evaluation exercise experiences of students and teachers. Tzu Chi Med J. 28(3): 113–120.

Yeates P, O'Neill P, Mann K, Eva K. 2013. Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. Adv Health Sci Educ. 18(3):325–341.

Yusuf L, Ahmed A, Yasmin R. 2018. Educational impact of mini-clinical evaluation exercise: a game changer. Pak J Med Sci. 34(2):405–411.