Adv in Health Sci Educ (2011) 16:131–142
DOI 10.1007/s10459-010-9235-6

REVIEW

In-training assessment using direct observation of single-patient encounters: a literature review

E. A. M. Pelgrim · A. W. M. Kramer · H. G. A. Mokkink · L. van den Elsen · R. P. T. M. Grol · C. P. M. van der Vleuten

Received: 1 October 2009 / Accepted: 12 May 2010 / Published online: 18 June 2010

Abstract We reviewed the literature on instruments for work-based assessment in single clinical encounters, such as the mini-clinical evaluation exercise (mini-CEX), and examined differences between these instruments in characteristics and feasibility, reliability, validity and educational effect. A PubMed search of the literature published before 8 January 2009 yielded 39 articles dealing with 18 different assessment instruments. One researcher extracted data on the characteristics of the instruments and two researchers extracted data on feasibility, reliability, validity and educational effect. Instruments are predominantly formative. Feasibility is generally deemed good and assessor training occurs sparsely but is considered crucial for successful implementation. Acceptable reliability can be achieved with 10 encounters. The validity of many instruments is not investigated, but the validity of the mini-CEX and the 'clinical evaluation exercise' is supported by strong and significant correlations with other valid assessment instruments. The evidence from the few studies on educational effects is not very convincing. The reports on clinical assessment instruments for single work-based encounters are generally positive, but supporting evidence is sparse. Feasibility of instruments seems to be good and reliability requires a minimum of 10 encounters, but no clear conclusions emerge on other aspects. Studies on assessor and learner training and studies examining effects beyond 'happiness data' are badly needed.

Keywords Educational effects · Feasibility · Mini-CEX · Reliability · Validity · Work-based assessment instruments

E. A. M. Pelgrim (✉) · A. W. M. Kramer · H. G. A. Mokkink · L. van den Elsen · R. P. T. M. Grol
Department of Primary Care and Community Care, Radboud University Nijmegen Medical Centre, Postbus 9101, Huispostnummer, 6500 HB Nijmegen, The Netherlands
e-mail: e.pelgrim@elg.umcn.nl

C. P. M. van der Vleuten
Maastricht University, Maastricht, The Netherlands

Introduction

The mini-clinical evaluation exercise (mini-CEX) is widely used for assessment in single work-based encounters of clinical competence at the top of Miller's pyramid—the 'does' level. Currently, assessment of clinical competence is receiving increasing attention, particularly in postgraduate training (Wass et al. 2001), and assessment of authentic performance is considered the main challenge. Reliable and valid performance measurements that can serve as a gold standard for clinical assessment have as yet not been achieved (Wass et al. 2001). Developed for the evaluation of a multitude of clinical competencies (Norcini et al. 1995), the mini-CEX is a single-encounter instrument to be used by professionals in conducting work-based assessment of actual clinical performance. It was originally developed in 1995 in the USA for the evaluation of internal medicine residents' clinical skills (Norcini et al. 1995, 2003) and its principal characteristics are direct observation of real patient encounters, easy and instant use in day-to-day practice, applicability in a broad range of settings and immediate feedback to the learner after the encounter. These characteristics make the mini-CEX an educational tool that can help learners to gain insight into the strengths and weaknesses of their clinical performance. It can be used to assess multiple competencies, such as communication and professionalism. Typically, the mini-CEX and similar instruments use global assessment scales, provide space for narrative comments and allow for feedback presented by a moderator in a post-encounter review session.

Since the mini-CEX was first introduced, several comparable instruments have been developed for use in undergraduate and postgraduate medical education, including, among many others, 'longitudinal evaluation of performance' (Prescott et al. 2002), 'structured clinical observation' (Lane and Gottlieb 2000) and the 'clinical encounter card' (Paukert et al. 2002). To our knowledge, however, no review has compared the characteristics and key qualities of these instruments. Feasibility, reliability, validity and educational effects are the core elements in determining the utility of assessment methods (van der Vleuten 1996). The only review of the validity of instruments for work-based clinical assessment was published in October 2009 (Kogan et al. 2009). The authors conclude that many tools are available, but evidence on their validity and descriptions of educational outcomes are scarce. We reviewed the literature on instruments for single-encounter work-based clinical assessment, like the mini-CEX. These instruments appear to hold promise for clinical assessment but too little is known about their characteristics and feasibility, reliability, validity and educational effects.

We addressed the following research questions:

1. What are the similarities and differences between the characteristics of clinical assessment instruments, such as the mini-CEX?
2. What is known about the feasibility, validity, reliability and educational effects of these clinical assessment instruments?

Methods

We conducted two searches of the PubMed database for papers on clinical assessment instruments published before 8 January 2009. For our first search, aimed at identifying papers dealing with the principal characteristics of work-based assessment instruments, we used the following search terms:

- clinical competence (medical subject heading [MeSH] term and text word) OR educational measurement (MeSH term and text word) OR educational measurements (text word) OR clinical skills (text word)

AND

- medical students (MeSH term and text word) OR clinical clerkship (MeSH term and text word) OR internship and residency (MeSH terms) OR internship (text word) OR residency (text word) OR medical education (MeSH term and text word) OR preceptorship (MeSH term and text word)

AND

- observation (text word) OR observe (text word) OR observed (text word)

AND

- feedback (MeSH term and text word)

OR

- reproducibility of results OR feasibility studies OR observer variation OR pilot projects OR psychometrics OR qualitative research OR statistical data interpretation OR Delphi-technique (MeSH terms and text words) OR evaluation studies (MeSH term, publication type and text word) OR validation studies (publication type and text word).

Our second PubMed search was limited to articles published between November 1995 (publication date of the first paper on the mini-CEX) and 8 January 2009, and used the text words:

- mini clinical evaluation exercise OR mini-CEX OR mCEX OR clinical evaluation exercise.

In addition, we manually searched the reference lists of the included articles for relevant articles.

We used the following inclusion criteria.

- the instrument is used by professionals to assess directly observed performance.
- the instrument is used in authentic patient encounters
- the instrument uses a generic and global assessment scale
- the instrument allows for feedback immediately after the assessment
- the instrument is used in a postgraduate or undergraduate medical programme.

And we applied the following exclusion criteria:

- the instrument is used for peer-, patient- or self-assessment
- the instrument only assesses technical skills
- the instrument is used in simulated encounters (as opposed to authentic encounters)
- the instrument (only) assesses a 'long case' (Wass and van der Vleuten 2004)
- the instrument reports results as a letter or comment
- no abstract is available

Articles were selected by four researchers (LvdE, EP, AK and HM). In an initial selection round, two researchers independently selected articles based on the title only. Any disagreements were resolved by discussion. Next, the abstracts of the articles selected in the first round were independently judged by two researchers. Any disagreements on

inclusion or exclusion were resolved in a meeting of three researchers. In the final selection round the full text of the remaining articles was read by LvdE or EP.

Data extraction

Data relating to the following characteristics of the assessment instruments were extracted from each article by one researcher (LvdE or EP).

- setting, summative or formative assessment
- type of encounters (e.g. in-patient, out-patient), assessor and learner
- subject of assessment
- rating scale, criteria for the allocation of marks, frame of reference
- the assessment form
- type of feedback (quantitative/qualitative)
- assessor training
- learner instruction.

Next, two of four researchers (LvdE, EP, AK and HM) extracted data related to the aspects addressed by the second research question:

- feasibility
- reliability
- validity
- educational effect.

Two of four researchers (LvdE, EP, AK and HM) analyzed each article to determine whether these four aspects were evaluated, which research methods were used and the outcomes of the study. If there was disagreement, a third or fourth researcher also read the article and consensus was reached through discussion. The data are presented in tables (appendices 1–6) that are available on https://www.umcn.nl/Onderwijs/IWOO/VOHA/Pages/OnderzoekbijdeVOHA.aspx. If an instrument was the subject of more than one article, additional articles were only included if they contained new information about the aspects of interest. Based on the tables, the researchers identified highlights and interesting results for each characteristic, which are reported in the results section.

Results

Descriptive analysis

The initial search yielded 349 articles. Of these, 261 were excluded based on the title, a further 50 were excluded based on the abstract and another 19 were eliminated after the reading of the full article. This left a total of 19 articles. The second search yielded 34 articles. After exclusion of five, nine and five articles based on title, abstract and full text, respectively, 15 articles from the second search met the criteria. The manual search of reference lists yielded another 5 articles. The resulting 39 articles dealt with 18 different assessment instruments (Alves de Lima et al. 2007; Anderson et al. 2005; Burch et al. 2006; Cook et al. 2008; Cook and Beckman 2009; Cruess et al. 2006; Donato et al. 2008; Dowson and Hassell 2006; Durning et al. 2002; Golnik et al. 2004; Golnik and Goldenhar 2005; Han et al. 2005; Hatala et al. 2006; Hatala and Norman 1999; Holmboe et al. 2003;

Kogan et al. 2002, 2003; Kogan and Hauer 2006; Lane and Gottlieb 2000; Links et al. 1984; Malhotra et al. 2008; Margolis et al. 2006; Nair et al. 2008; Norcini et al. 1995, 1997, 2003; Norcini and Burch 2007; Nyman and Sheridan 1997; Paukert et al. 2002; Prescott et al. 2002; Prescott-Clements et al. 2008; Richards et al. 2007; Ringsted et al. 2003; Ross 2002; Shayne et al. 2002, 2006; Torre et al. 2007; Turnbull et al. 2000; Wilkinson et al. 2008), which are listed in Table 1.

Characteristics

The instruments included in the review assess a wide range of competencies or combinations of competencies. Some allow coverage of broad content and can be used in all kinds of clinical situations; others assess content that is limited to a particular setting, e.g. a palliative care or psychiatry clerkship. All instruments itemize content globally, but some are more detailed than others (items such as: 'open-ended questions' versus 'patient communication'). Most items relate to the 'medical expertise', 'communication' and 'professionalism' competencies from the Canadian Medical Educational Directives for Specialists (CanMEDS). Some items relate to the CanMEDS competence 'management skills'. Generally, the instruments appear to be flexible with regard to content. They can be used to assess a multitude of competencies and are easily attuned to a specific educational context.

Most instruments are (intended to be) used for formative purposes. It is consistent with this purpose that almost all instruments ask for qualitative, narrative feedback to be provided in writing or orally. Additionally, almost all instruments require quantitative feedback on a rating scale. These scales vary widely, ranging from dichotomized scores of

Table 1 Assessment instruments

1	Mini clinical evaluation exercise
2	Ophthalmic clinical evaluation exercise
3	Palliative care clinical evaluation exercise
4	Professionalism mini evaluation exercise
5	Competence based assessment, rheumathology
6	Structured clinical observation
7	Patient evaluation assessment form
8	Global rating form in anaesthesiology
9	Ward rating form (in clinical work sampling approach to in-training assessment)
10	Clinical-performance biopsy instrument
11	Clinical evaluation exercise (in emergency medicine training programme)
12	Clinical skills assessment form, direct observation exercise
13	Standardized direct observation assessment tool
14	Evaluation of consulting skills (of trainee general practitioners)
15	Longitudinal evaluation of performance
16	Minicard
17	Clinical encounter card
18	Bedside formative assessment

'satisfactory' and 'unsatisfactory' to an 11-point scale. A minority of the instruments (four with small and three with large scales) provide criteria for the allocation of marks or behavioural anchors. Only one study examines the effects of different rating scales (Cook and Beckman 2009) by comparing the results for 9- and 5-point scales. Inter-rater reliability was similar for both scales, but the 9-point scale showed better agreement with previously established levels of competence of a performance on video (the scripted competence level). Based on the assumption that previously established competence levels are accurate, the 9-point scale was better able to accurately classify learners' competence as unsatisfactory or superior. A reference norm for competence rating is specified in no more than eight instruments: five use an 'end of training' norm and three a 'class level' norm. However, norm selection is not based on evidence and authors generally state few or no arguments to support their choice of rating scale or frame of reference, thereby leaving much freedom of interpretation to assessors.

Assessors almost always receive some form of training before an instrument is implemented. Training involves verbal instruction or a workshop, but it is uncommon for training effects to be evaluated. The only authors to do so are Cook et al. (2008), who evaluated the effects of a workshop on error training, performance dimension training, behavioural observation training and frame of reference training using lecture, video and facilitated discussion. They found no improvement in inter-rater reliability of mini-CEX scores in a group of assessors who had attended the workshop compared to a control group receiving no training. Generally, learner instruction receives scant attention. If learners are instructed at all they receive verbal or written instructions, but no studies evaluate the effects.

In conclusion, instruments show considerable variation in content, rating scale, frame of reference, assessor training and learner instruction. There is a striking paucity of research on these characteristics, which are merely described in the majority of studies without evidence to support their value.

Feasibility

Studies of feasibility mostly focus on completion rates of the instruments or users' satisfaction. Feasibility is generally qualified as good but no clear criteria are set in advance and results vary. Durning et al. (2002) and Torre et al. (2007), for example, report completion rates of 96.4 and 100%, respectively, but Turnbull et al. (2000) conclude that feasibility is good with a response rate of only 23%.

Conclusions regarding the feasibility of the various instruments, with the exception of the mini-CEX, are based on single studies. When more studies are available, the results are both negative and positive. Wilkinson et al. (2008) attribute feasibility problems to lack of time and the fact that the procedure is experienced as time consuming. Alves de Lima et al. (2007) blame poor feasibility on inadequate implementation. They conclude that assessment instruments must be well integrated within the curriculum and part of the routine of practice, and additionally propose that workshops are a better way to implement an instrument than written instructions. Clearly, further studies are needed to unravel the instruments' feasibility issues.

Reliability

Generalizability or reproducibility was studied for four instruments in eight studies. The results are presented in Table 2. We used the Spearman-Brown formula to calculate the average reliability coefficient for all instruments. For most of them acceptable reliability

Table 2 Generalizability analysis

Reference no.	Instrument	Raters	Encounters	Reliability coefficient	Reliability with 8 encounters (Spearman Brown formula)	Reliability with 10 encounters (Spearman Brown formula)	Reliability with 12 encounters (Spearman Brown formula)
Waas et al. (2001)	Mini-CEX	8	16	≥0.80	≥0.25	≥0.71	≥0.75
Margolis et al. (2006)	Mini-CEX	1	10	0.39	0.34	0.39	0.43
Margolis et al. (2006)	Mini-CEX	10	10	0.83	0.57	0.83	0.85
Nair et al. (2008)	Mini-CEX	1	8	0.88	0.88	0.90	0.92
Alves de Lima et al. (2007)	Mini-CEX			10 evaluations for a minimally reliable inference		Reliable	Reliable
Kogan et al. (2003)	Mini-CEX	4	Probably 4	0.62	0.77	0.80	0.83
Kogan et al. (2003)	Mini-CEX	6	Probably 6	0.71	0.77	0.80	0.83
Kogan et al. (2003)	Mini-CEX	8	Probably 8	0.77	0.77	0.81	0.83
Cruess et al. (2006)	P-MEX	Probably 1	10 (a 12)	≥0.80	≥0.76	≥0.80	≥0.83
Turnbull et al. (2000)	WRF	Probably 1	3.2 forms completed	≥0.70	≥0.85	≥0.88	≥0.90
Richards et al. (2007)	CEC	7	20	0.58	0.36	0.41	0.45
Richards et al. (2007)	CEC	12	18	0.69	0.49	0.55	0.60
Total					0.59	0.69	0.73

(>0.8) can be achieved with a sample of 10 encounters. In other words, reliability seems achievable with a feasible sample of encounters. For some studies, we could not determine the number of assessors involved. The study by Margolis et al. (2006) is the only one to examine reliability with different numbers of assessors. The results show that one assessor taking 10 encounters is much less reliable than 10 assessors taking one encounter each (0.39 and 0.83, respectively). This outcome is contradicted by Nair et al. (2008), who conclude that the mini-CEX is reliable (0.88) with one assessor and eight encounters. However, this study did not explicitly examine the effects of different numbers of assessors. More research is needed to systematically tease out sources of variance in reliability to enable well founded recommendations with regard to the required numbers of (different) assessors and encounters.

Ringsted et al. (2003) explain the low inter-rater reliability of their 'global rating form in anaesthesiology' by staff being unfamiliar with the instrument's underlying concept. They suggest that intensive assessor training might improve reliability results, but the opposite conclusion is put forward by Cook et al. (2008). This conflicting evidence underlines the need for more research into inter-rater reliability and how it is affected by assessor training.

Validity

Criterion validity of the mini-CEX and the 'clinical encounter card' was evaluated by comparisons of the results with those of instruments of proven validity. For the mini-CEX, strong and significant correlations were found with results on the Royal College of Physicians and Surgeons of Canada Comprehensive Examination in Internal Medicine (RCSPC-IM), a high-stakes assessment of clinical competence. Correlations were 0.73 with the subscale 'structured oral', 0.67 with the subscale 'bedside station' and 0.72 with the subscale 'written examination' (Hatala et al. 2006). In addition, strong correlations are reported between mini-CEX scores and corresponding scores on a monthly evaluation form and 'in-training examination scores' (Durning et al. 2002). The 'clinical encounter card' showed significant positive correlations with learners' 'clinical performance ratings', 'final grades' and scores on an important summative examination (National Board of Medical Examiners [NBME]) (Richards et al. 2007). Interestingly, no correlations are reported between the 'clinical encounter card' and an objective structured clinical examination (OSCE).

A number of studies infer construct validity from an increase in ratings over time. Kogan et al. (2003) report an increase in mean scores on the mini-CEX during one year. Links et al. (1984) found significant improvement in skills as manifested in pre- and post-observations, using the 'clinical skills assessment form'. Prescott-Clements et al. (2008) report improvement in ratings on 'longitudinal evaluation of performance' in the course of 1 year.

In conclusion, the validity of the mini-CEX and the 'clinical encounter card' appears to be supported by strong and significant correlations with other assessment instruments. For some other instruments positive indications for construct validity are reported, but for most instruments evidence of validity remains to be provided.

Educational effect

Some studies evaluated educational effect by eliciting learners' or assessors' attitudes towards the use of the instrument, but none of the studies examined educational effects by measuring improvement of clinical skills or the quality of patient care. Although authors emphasize the formative nature of assessment procedures, they examine effects on learning

and performance by evaluating users' subjective judgements or perceived satisfaction. For the most part, the reported effects are positive. Learners rated the value of 'structured clinical observation' four on a five-point scale (Lane and Gottlieb 2000) and rated the 'clinical skills assessment form' as the second most valuable component of their clerkship in terms of assisting skill acquisition (Links et al. 1984). Outcomes of a student questionnaire on 'bedside formative assessment' show that 95.6% recognize its learning value, 70% acknowledge the informative, advisory and motivational role of feedback and 71.9% report that the assessment stimulated them to do more preparatory reading.

However, outcomes like learning behaviour, transfer of skills to new situations or improvement of patient care are not investigated, although they are crucial for the evaluation of educational impact. Currently, educational effects are a neglected area of assessment research, which should be given much greater priority in future research.

Discussion

As for the similarities and differences between the characteristics of the instruments, the main conclusion is that there is huge variation in the competencies being assessed, rating scales, frame of reference, assessor training and learner instruction. Unfortunately, there is hardly any sound research reported on these characteristics. Authors describe rating scales, frames of reference and assessor training but fail to elaborate on rationales and usually do not investigate their value. Consequently, assessment characteristics remain implicit and interpretation is largely left to assessors. This will inevitably have a profound effect on instruments' measurement characteristics.

Almost all the instruments discussed in this review originated after the introduction of the mini-CEX at the Medical College of Pennsylvania, Philadelphia. An exception is the 'clinical skills assessment form', an observation exercise that was introduced in the psychiatric clerkship at McMaster University, Canada as early as 1984 (Links et al. 1984), well before the publication of the first paper on the mini-CEX. It is interesting to note that this early appearance on the medical education scene of a predecessor of the mini-CEX apparently failed to make much of an impact either in the literature or in educational practice. Perhaps the time was not ripe then for this type of instrument.

Some information on the feasibility, validity, reliability and educational effect of the instruments we studied emerges from the review. Conclusions regarding feasibility are generally positive. Despite the absence of direct compelling evidence, we are inclined to conclude that training may be the key to effective implementation of instruments because it can improve the quality of their use. The value of these instruments lies mainly in the process of formative feedback and thus in the feedback skills of assessors and the extent to which they pay serious attention to this process. Much of what is assessed is left implicit and is up to the discretion of assessors (Holmboe et al. 2003; Norcini and Burch 2007). Assessors need training to reliably rate learners' performance and discriminate between performance levels (Kogan et al. 2009). For learners too training may play an important role, although no direct evidence is available to support this. It seems likely that learner training can increase feasibility and educational effect.

Criterion validity was only evaluated for the mini-CEX and the 'clinical encounter card', and these instruments showed strong and significant correlations with other assessment instruments. Construct validity was inferred from three studies showing that ratings increased over time. Otherwise, like Kogan et al.'s review of validity (Kogan et al. 2009), our review reveals a general lack of evidence of validity.

The outcomes of reliability studies suggest that around 10 encounters suffice for a reproducible outcome. This is somewhat surprising. In terms of testing time (time of one medical consultation) 10 encounters compares favourably with the samples needed for other standardized and objectified assessment formats (Van der Vleuten and Schuwirth 2005), although one would expect poor reliability of an instrument characterized by absence of explicit characteristics. Apparently (different) assessors pick up measurement information that is relatively generalizable across individual encounters, while at the same time broad sampling across assessors evens out the effects of assessor subjectivity. Good reliability is no guarantee for the absence of bias, however, and, due to their quite subjective nature, instruments like the mini-CEX may actually be quite vulnerable to bias. All this requires further investigation. We also need more evidence regarding factors that contribute to (un)reliability and the extent of this contribution to underpin recommendations on sound sampling strategies.

Evidence on educational effect is lacking as well. No studies examined whether instruments improve learning, clinical skills or the quality of patient care. Given the formative nature of the instruments, effects on learning and performance are more or less the prime objective of this type of assessment. Existing research typically evaluates perceptions of users, and although the outcomes are overwhelmingly positive, they do not provide compelling evidence for learning effects. More rigorous research will have to elucidate the educational effects of clinical work-based assessment.

An important conclusion from our review appears to be that instruments for authentic work-based assessment of single clinical encounters should not be evaluated outside the context of the curriculum or other assessment instruments. Assessment by one instrument can only be a part of the whole story. The 'competence based assessment, rheumatology' for example was not valid when applied in isolation (Dowson and Hassell 2006). It should be used as a component of a spectrum of assessment instruments that complement each other. While optimization of the feasibility, validity, reliability and educational effect of individual instruments is important, it is equally, if not more, important to look from a broader perspective at the respective unique contributions of different instruments to the assessment of clinical competence (Van der Vleuten and Schuwirth 2005). Assessment procedures should be integrated within the curriculum and preferably also be an integral part of routine practice (Alves de Lima et al. 2007).

It should be noted that we included articles in the review on the basis of the subjects they addressed, not the quality of their research. Some bias may have arisen because we did not systematically judge research quality.

In so far as the articles report on feasibility, validity, reliability or educational effect, the conclusions are mostly positive. This absence of negative or critical outcomes could be suggestive of publication bias. It cannot be ruled out that studies on inadequate instruments were not published.

Although single-encounter clinical assessment instruments appear to be received positively in the literature, this positive reception is based on relatively limited empirical justification. Results on the most extensively evaluated aspects, feasibility and reliability, support the viability of the format and the use of a minimum of 10 encounters to attain reliability. However, there is an obvious need for further, and especially more scientifically rigorous, research on all the characteristics that we studied. We also need further research on basic characteristics like rating scales, narrative feedback, frame of reference, etc. Although a call for more and better research may be the sad conclusion from most reviews, it is unfortunately equally applicable to single-encounter work-based clinical assessment instruments.

References

Alves de Lima, A., Barrero, C., Baratta, S., Castillo Costa, Y., Bortman, G., Carabajales, J., et al. (2007). Validity, reliability, feasibility and satisfaction of the mini-clinial evaluation exercise (mini-CEX) for cardiology residency training. *Medical Teacher, 29*, 785–790.

Anderson, C. I., Jentz, A. B., Harkema, J. M., Kareti, L. R., Apelgren, K. N., & Slomski, C. A. (2005). Assessing the competencies in general surgery residency training. *The Americal Journal of Surgery, 189*, 288–292.

Burch, V. C., Seggie, J. L., & Gary, N. E. (2006). Formative assessment promotes learning in undergraduate clinical clerkships. *South African Medical Journal, 96*, 430–433.

Cook, D. A., & Beckman, T. J. (2009). Does scale length matter? A comparison of nine-versus five-point rating scales for the mini-CEX. *Advances in Health Sciences Education, 14*, 655–664.

Cook, D. A., Dupras, D. M., Beckman, T. J., Thomas, K. G., & Pankratz, S. (2008). Effect of rater training on reliability and accuracy of mini-CEX scores: A randomized, controlled trial. *Journal of General Internal Medicine, 24*, 74–79.

Cruess, R., McIlroy, J. H., Cruess, S., Ginsburg, S., & Steinert, Y. (2006). The professionalism mini-evaluation exercise: A preliminary investigation. *Academic Medicine, 81*, S74–S78.

Donato, A. A., Pangaro, L., Smith, C., Rencic, J., Diaz, Y., Mensinger, J., et al. (2008). Evaluation of a novel assessment form for observing medical residents: A randomised, controlled trail. *Medical Education, 42*, 1234–1242.

Dowson, C., & Hassell, A. (2006). Competence-based assessment of specialist registrars: Evaluation of a new assessment of out-patient consultations. *Rheumatology, 45*, 459–464.

Durning, S. J., Cation, L. J., Markert, R. J., & Pangaro, L. N. (2002). Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. *Academic Medicine, 77*, 900–904.

Golnik, K. C., & Goldenhar, L. (2005). The ophthalmic clinical evaluation exercise: Reliability determination. *Opthalmology, 112*, 1649–1654.

Golnik, K. C., Goldenhar, L. M., Gittinger, J. W., & Lustbader, J. M. (2004). The ophthalmic clinical evaluation exercise (OCEX). *Ophthalmology, 111*, 1271–1274.

Han, P. K., Keranen, L. B., Lescisin, D. A., & Arnold, R. M. (2005). The palliative care clinical evaluation exercise (CEX): An experience-based intervention for teaching end-of-life communication skills. *Academic Medicine, 80*, 669–676.

Hatala, R., Ainslie, M., Kassen, B. O., Mackie, I., & Roberts, J. M. (2006). Assessing the mini-clinical evaluation exercise in comparison to a national specialty examination. *Medical Education, 40*, 950–956.

Hatala, R., & Norman, G. R. (1999). In-training evaluation during an internal medicine clerkship. *Academic Medicine, 74*, S118–S120.

Holmboe, E. S., Huot, S., Chung, J., Norcini, J., & Hawkins, R. E. (2003). Construct validity of the mini clinical evaluation exercise (miniCEX). *Academic Medicine, 78*, 826–830.

Kogan, J. R., Bellini, L. M., & Shea, J. A. (2002). Implementation of the mini-CEX to evaluate medical students' clinical skills. *Academic Medicine, 77*, 1156–1157.

Kogan, J. R., Bellini, L. M., & Shea, J. A. (2003). Feasibility, reliability, and validity of the mini-clinical evaluation exercise (mCEX) in a medicine core clerkship. *Academic Medicine, 78*, S33–S35.

Kogan, J. R., & Hauer, K. E. (2006). Brief report: Use of the mini-clinical evaluation exercise in internal medicine core clerkships. *Journal of General Internal Medicine, 21*, 501–502.

Kogan, J. R., Holmboe, E. S., & Hauer, K. E. (2009). Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. *JAMA, 302,12*, 1316–1326.

Lane, J. L., & Gottlieb, R. P. (2000). Structured clinical observations: A method to teach clinical skills with limited time and financial resources. *Pediatrics, 105*, 973–977.

Links, P. S., Colton, T., & Norman, G. R. (1984). Evaluating a direct observation exercise in a psychiatric clerkship. *Medical Education, 18*, 46–51.

Malhotra, S., Hatala, R., & Courneya, C. A. (2008). Internal medicine residents' perceptions of the mini-clinical evaluation exercise. *Medical Teacher, 30*, 414–419.

Margolis, M. J., Clauser, B. E., Cuddy, M. M., Ciccone, A., Mee, J., Harik, P., et al. (2006). Use of the mini-clinical evaluation exercise to rate examinee performance on a multiple-station clinical skills examination: A validity study. *Academic Medicine, 81*, S56–S60.

Nair, B. R., Alexander, H. G., McGrath, B. P., Parvathy, M. S., Kilsby, E. C., Wenzel, J., et al. (2008). The mini clinical evaluation exercise (mini-CEX) for assessing clinical performance of international medical graduates. *The Medical Journal of Australia, 189*, 159–161.

Norcini, J. J., Blank, L. L., Arnold, G. K., & Kimball, H. R. (1995). The mini-CEX (clinical evaluation exercise): A preliminary investigation. *Annals of Internal Medicine, 123*, 795–799.

Norcini, J. J., Blank, L. L., Arnold, G. K., & Kimball, H. R. (1997). Examiner differences in the mini-CEX. *Advances in Health Sciences Education, 2*, 27–33.

Norcini, J. J., Blank, L. L., Duffy, F. D., & Fortna, G. S. (2003). The mini-CEX: A method for assessing clinical skills. *Annals of Internal Medicine, 138*, 476–481.

Norcini, J., & Burch, V. (2007). Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Medical Teacher, 29*, 855–871.

Nyman, K. C., & Sheridan, B. (1997). Evaluation of consulting skills of trainee general practitioners. *Australian Family Physician, 26*, S28–S35.

Paukert, J. L., Richards, M. L., & Olney, C. (2002). An encounter card system for increasing feedback to students. *The American Journal of Surgery, 183*, 300–304.

Prescott, L. E., Norcini, J. J., McKinlay, P., & Rennie, J. S. (2002). Facing the challenges of competency-based assessment of postgraduate dental training: Longitudinal evaluation of performance (LEP). *Medical Education, 36*, 92–97.

Prescott-Clements, L., van der Vleuten, C. P. M., Schuwirth, L. W. T., Hurst, Y., & Rennie, J. S. (2008). Evidence for validity within workplace assessment: The longitudinal evaluation of performance (LEP). *Medical Education, 42*, 488–495.

Richards, M. L., Paukert, J. L., Downing, S. M., & Bordage, G. (2007). Reliability and usefulness of clinical encounter cards for a third-year surgical clerkship. *Journal of Surgical Research, 140*, 139–148.

Ringsted, C., Ostergaard, D., Ravn, L., Pedersen, J. A., Berlac, P. A., & van der Vleuten, C. P. M. (2003). A feasibility study comparing checklists and global rating forms to assess resident performance in clinical skills. *Medical Teacher, 25*, 654–658.

Ross, R. (2002). A clinical-performance biopsy instrument. *Academic Medicine, 77*, 268.

Shayne, P., Gallahue, F., Rinnert, S., Anderson, C. L., Hern, G., & Katz, E. (2006). Reliability of a core competency checklist assessment in the emergency department: The standardized direct observation assessment tool. *Academic Emergency Medicine, 13*, 727–732.

Shayne, P., Heilpern, K., Ander, D., & Palmer-Smith, V. (2002). Protected clinical teaching time and a bedside clinical evaluation instrument in an emergency medicine training program. *Academic Emergency Medicine, 9*, 1342–1349.

Torre, D. M., Simpson, D. E., Elnicki, D. M., Sebastian, J. L., & Holmboe, E. S. (2007). Feasibility, reliability and user satisfaction with a PDA-based mini-CEX to evaluate the clinical skills of third-year medical students. *Teaching and Learning in Medicine, 19*, 271–277.

Turnbull, J., MacFadyen, J., van Barneveld, C., & Norman, G. (2000). Clinical work sampling. A new approach to the problem of in-training evaluation. *Journal of General Internal Medicine, 15*, 556–561.

van der Vleuten, C. P. M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education, 1*, 41–67.

van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*, 309–317.

Wass, V., & van der Vleuten, C. (2004). The long case. *Medical Education, 38*, 1176–1180.

Wass, V., van der Vleuten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. *Lancet, 357*, 945–949.

Wilkinson, J. R., Crossley, J. G. M., Wragg, A., Mills, P., Cowan, G., & Wade, W. (2008). Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Medical Education, 42*, 364–373.