Assessment of clinical competence: Reliability, Validity, Feasibility and Educational Impact of the mini-CEX

Alberto Alves de Lima
Assessment of clinical competence: Reliability, Validity, Feasibility and Educational Impact of the mini-CEX

DISSertation

to obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus Prof. Dr. L.L.G. Soete,
in accordance with the decision of the Board of Deans,
to be defended in public on Wednesday 23rd October 2013 at 16.00 hrs.

by

Alberto Alves de Lima
Promotor
Prof. dr. C.P.M. van der Vleuten

Beoordelingscommissie
Prof. dr. A.J.J.A. Scherpber (chair)
Dr. E.W. Driessen
Prof. dr. S. Heeneman
Dr. A. Kramer, UMC St. Radboud Nijmegen
Prof. dr. F. Scheele, VUMC Amsterdam
Acknowledgments

I would like to express my gratitude to the Board of Directors and to the Department of Education and Research of the Instituto Cardiovascular de Buenos Aires.

To my promotor: Professor Cees Van der Vleuten for his help and support.

My last and most profound debt of gratitude goes to my wife Ingrid and to my two children, Tomas and Agustina, without whom this project would not have been possible.
## Contents

Chapter 1 ................................................................................................................. 9
**General introduction**

Chapter 2 ........................................................................................................... 19
**Validity, reliability, feasibility and satisfaction of the Mini-Clinical Evaluation Exercise (Mini-CEX) for cardiology residency training**
*Medical Teacher* 2007; 29:785-790

Chapter 3 ........................................................................................................... 31
**A qualitative study of the impact on learning of the mini clinical evaluation exercise in postgraduate training**
*Medical Teacher* 2007; 29:785-790

Chapter 4 ........................................................................................................... 45
**Teachers' experiences of the role and function of the mini clinical evaluation exercise in post-graduate training**
*International Journal of Medical Education* 2010; 1:68-73

Chapter 5 ........................................................................................................... 59
**A laboratory study on the reliability estimations of the Mini-CEX**
*Advances Science Health Education* 2013; 18 (1):5-13

Chapter 6 ........................................................................................................... 71
**Educational impact of the mini-CEX on medical students' performance**
*This paper was accepted in a resumed version in Medicina (Bs As), 2013, 73(3): 286-291*

Chapter 7 ........................................................................................................... 79
**Understanding raters errors: Reliability estimations of the mini-CEX using traditional and construct-aligned scales**
*Submitted*

Chapter 8 ........................................................................................................... 91
**Discussion**
Chapter 1

General introduction
Introduction

Historically, assessment has often been directed to what is measurable rather than to what is important. Over the last 30 years, however, we have witnessed a gradual shift of focus toward teaching and assessing what matters most (Crossley & Jolly, 2012). Firstly, this was reflected in moving from testing superficial factual knowledge towards testing of understanding the material (Biggs & Collis, 1982). Secondly, in recognizing that skills and attitudes are as important as knowledge (Bloom, 1956). Finally, the psychometric perspective highlighted assessors’ subjectivity and case specificity of performance and prompted the movement towards multiple mini-test samplings across different assessment formats, such as the mini-CEX (Norcini, Blank, Duffy, & Fortna, 2003). Many of these developments have deconstructed assessment and deconstructed learning, as some would argue. In other words, breaking up behaviour into assessable subcomponents has directed learners’ attention to discrete elements of competencies instead of focusing on the big picture. Interestingly, the competency movement advocates moving to the opposite direction by arguing that, in practice, all the components should be integrated to achieve competency in dealing with real problems (McClelland, 1973). Miller’s pyramid reflects these ideas by implying that knowledge is crucial but not sufficient for understanding (Miller, 1990), understanding is crucial but not sufficient to demonstrate ability, and ability demonstrated in a controlled setting is not sufficient to imply competency in day-to-day real practice. In other words, the consecutive levels or layers of the pyramid reconstruct what has been separated (Crossley & Jolly, 2012). This has proved to be highly relevant in assessment of competency in studies demonstrating that doctors’ abilities as assessed in a controlled environment do not dependably predict their performance in day-to-day practice (Rethans, Sturman, Drop, van der Vleuten, & Hobus, 1991). In order to obtain a reliable impression of doctors’ performance in routine practice situations, assessment should be targeted at doctors while engaged in their normal work. This type of assessment seems to rely crucially on observation of performance in the workplace, and observation is notoriously rare in clinical assessment. A study by Day et al. (1990) in the United States documented that the vast majority of first-year trainees in internal medicine were not observed more than once by a faculty member when taking a history or doing a physical examination during a patient encounter (Day et al., 1990). These findings have been reported in numerous studies, and the paucity of observation of performance is regrettable not only because it precludes proper assessment of basic clinical skills but also because it deprives learners of essential feedback on performance to guide their performance improvement.

To remedy these shortcomings by encouraging observation of performance by faculty, the American Board of Internal Medicine proposed the use of the mini clinical evaluation exercise (mini-CEX) (Norcini, Blank, Arnold, & Kimball, 1995). An
advantage of the mini-CEX and other workplace-based methods is that they fulfil the three basic requirements for assessment techniques that facilitate learning: 1) alignment of the content of the training program, the competencies expected as outcomes, and assessment practices; 2) feedback to trainees during and/or after assessment events; 3) strategic usage of assessment events to steer trainee learning toward the desired outcomes. The mini-CEX is a valuable tool to evaluate clinical competence because it necessitates direct observation of a learner engaged in a clinical encounter, rating of performance for a set of competencies, and a feedback session immediately after observed performance.

Before going into the details of the instrument, we will discuss the three key components of the mini-CEX: clinical competence, direct observation, and feedback.

**Clinical competence**

Clinical competence is defined as the degree to which individuals can use their knowledge, skills, and attitudes in an integrative way to successfully carry out complex professional tasks in their daily practice (Kane, 1992). Clinical competence is multidimensional, that is, during patient encounters doctors need to integrate and perform different task components, such as communication and physical examination.

Because competent clinical performance is highly context dependent, competence demonstrated in one particular case scenario (for example: in the coronary care unit) does not automatically guarantee successful performance in another scenario. This is known as the problem of content specificity of clinical performance, and it implies that general conclusions about a learner’s competence must be based on many assessments conducted across different contexts, settings, and cases.

Miller’s simple conceptual model of clinical performance clearly illustrates what medical educators can measure in terms of levels of competence (Miller, 1990). Miller visualizes competence as a pyramid with a base of factual knowledge, i.e. the ‘knows’ level, and three further levels. One level up from ‘knows’ is the level of ‘knows how’, concerned with how to use knowledge in a particular context, which is a close approximation of the concepts of clinical reasoning and problem solving. One level higher, the ‘shows how’ level reflects the ability to act in a practical situation showing appropriate and competent hands-on behaviour in simulated practice situations. Finally, the ‘does’ level refers to authentic performance in realistic workplace situations. As one moves up in the pyramid, assessment needs to be more clinically authentic. If examiners wish to assess performance at the highest level of Miller’s pyramid, they have no choice but to observe habitual performance in everyday practice (van der Vleuten, 2000).
Direct observation

Assessment of clinical competencies based on direct observation of a learner interacting with a patient constitutes a tool of inestimable learning value. Regular direct observation facilitates longitudinal follow-up of a learner’s progress on the one hand while, on the other hand, it enables ‘in situ’ and ‘in vivo’ correction and reinforcement of a learner’s actions and attitudes (Holmboe, 2004). Ideally, the clinical supervisor gathers and records information about the observed situation, usually by means of a checklist or rating scale, but, in reality, observation unfortunately occurs too infrequently and documentation is not really adequate. End-of-rotation global rating forms are often completed by supervisors who have not directly observed the learner in encounters with patients. Nonetheless, assessment of competence displayed in authentic behaviour that is directly observed by the assessor is the crucial component of any clinical learning and a prerequisite for trustworthy and defensible certification of competencies resulting from that learning.

Constructive feedback

Constructive feedback is defined as the act of giving information to a learner by describing his/her performance in an observed clinical situation. The elements of feedback that are required to provide the learner effective guidance to improve performance are: observation of an event, appraisal of the event according to a standard, and recommendations for improvement (Alves de Lima, 2008). The impact of feedback is optimized when learners compare their teacher’s feedback with their self-assessment of the same performance. Dissonance between desired and actual performance can give strong motivational impetus issue to deep learning. The purpose of constructive feedback is to provide guidance and advice on how to enhance future performance in line with desired objectives (Carr, 2006; Ende, 1983).

The history of the mini-CEX

In 1972, the American Board of Internal Medicine (ABIM) adopted the Clinical Evaluation Exercise (CEX). The CEX consists of a bedside oral exam and is widely used in postgraduate training programs for end-of-year assessments or on completion of residency (Norcini et al., 1995). In its traditional format, the CEX is conducted by a clinical supervisor who observes a learner during a patient interview, while the learner carries out a physical examination on the patient, gives a presen-
tation of the findings, and delivers a proposed diagnostic-therapeutic strategy. At the end of the CEX the clinical supervisor provides feedback to the learner about his/her performance. The entire exercise takes two hours. As an assessment method, the traditional CEX presents three major problems. First, the learner is evaluated by a single clinical supervisor, which is questionable in view of the known dissonance among observers (Norcini et al., 2003). Second, the assessment is based on one single patient encounter. From the problem of content specificity we know that performance on one case is unlikely to predict performance on other cases and therefore not generalizable to other patients (Eva, 2003). Third, a single long case takes too much time, which reduces feasibility.

**What is the mini-CEX?**

The mini-Clinical Evaluation Exercise (mini-CEX) focuses on the core skills that learners should demonstrate during patient encounters and requires teachers to document a learner’s performance in six general competencies. The mini-CEX is easy to use by clinical supervisors, because it fits seamlessly in the daily routine of any clinical setting. Estimated time of the interaction does not exceed 15-20 minutes, and a learner should annually receive at least eight evaluations from different clinical supervisors. One single clinical supervisor observes and evaluates a learner taking a focused history and performing a physical examination. After the learner has presented the diagnostic and treatment plan, the clinical supervisor completes a short evaluation form and gives feedback to the learner. As the encounter is relatively short and takes place as a natural part of routine practice within the training environment, it is quite feasible to have different supervisors evaluate learners on different cases and different occasions during the course of the residency program.

On the nine-point rating scale that is used, four is formally classified as satisfactory but actually denotes ‘marginal’ performance indicating that the learner needs to improve performance by engaging in recommended remediation to ensure that the requirements for board certification can be met. The competencies the mini-CEX developed by Norcini et al. (1995) can evaluate are defined as follows:

1) **Medical Interviewing Skills**: the learner facilitates the patient’s story-telling through effective use of questions/directions in order to obtain accurate and required information; the learner responds appropriately to affect and non-verbal cues.

2) **Physical Examination Skills**: the examination is conducted efficiently and in a logical sequence; the learner balances screening/diagnostic steps to map the patient’s problem, informs the patient, and is sensitive to the patient’s comfort and modesty.
3) **Humanistic Qualities/ Professionalism:** the learner shows respect, compassion, empathy, and establishes trust; the learner attends to the patient’s needs for comfort, modesty, confidentiality, and information.

4) **Clinical Judgment:** the learner selectively orders/perform appropriate diagnostic investigations/tests, and considers risks and benefits.

5) **Counselling Skills:** the learner explains the rationale for tests/treatment options, obtains the patient’s informed consent, educates/counsels the patient on the proposed management.

6) **Organization/ Efficiency Skills:** the learner prioritizes actions; uses time efficiently; is succinct.

7) **Overall Clinical Competence:** the learner demonstrates good judgment, synthesis, caring, effectiveness, and efficiency.

Performance is rated on a nine-point scale, with 1, 2, and 3 indicating unsatisfactory performance, 4 marginal performance, 5 and 6 satisfactory performance, and 7, 8, and 9 superior performance. In addition to the performance data, the clinical supervisor records information about the setting of the assessment, such as the inpatient service, the outpatient clinic, or the emergency department, the complexity of the case (low, moderate, high), and the patient’s gender, age, and major medical problems and diagnoses (Norcini et al., 1995).

**The problem statement**

In general, evaluation methods are biopsies of knowledge, skills, and attitudes believed to describe a defined trait that predicts a specific aspect of future performance (Turnbull & van Barneveld, 2002). In order to be meaningful, evaluation methods must reflect previously developed objectives. An evaluation tool to measure clinical competence may vary in appropriateness depending upon the target competency, and the appropriateness of the evaluation method depends on its utility. The utility of an assessment tool is defined as the degree to which it measures the desired objective (validity), the consistency or reproducibility of the scores (reliability), the educational benefit (educational impact), the acceptance by teachers and students (acceptability), and the cost-benefit ratio of information elicited and resources (costs) used (van der Vleuten, 1996). The relationship among the variables is multiplicative: if one of the elements is zero, the overall utility of the assessment tool will be zero. In general, the main problem related to methods for the assessment of clinical competency is the need for multiple and objective observations of performance in order to attain acceptable reliability of observation (Crossley & Jolly, 2012). If two similar assessment tools are administered to a group of learners, it is unlikely that each learner will receive exactly the same score on both instruments. Similarly, if two raters observe the performance of the same
learner, they are quite likely not to assign exactly the same grade. Each score assigned to a learner is subjected to measurement error, which may be random, happening purely by chance, or systematic, affecting scores in a consistent manner. Both random and systematic errors influence the interpretation of assessment scores. Thus, the extent to which we can minimize error variance or signal noise while maximizing true variance is an important aspect of any measurement tool. It is desirable to have a consistent (or reproducible) measurement of medical trainee clinical performances. The reliability of clinical assessment methods, i.e. their ability to consistently detect differences in performance, has been investigated in a variety of contexts. Potential sources of measurement errors identified in the literature include: rater stringency/leniency, student characteristics, rater-student interaction, the measurement tool/task, the occasion of observation, and the timeliness of the evaluation (Crossley & Jolly, 2002). Unfortunately, efforts to meet all these requirements inevitably have an adverse effect on the feasibility of the assessment. Since it is clearly inadvisable to use one single measurement, a number of measurements should be integrated into an overall comprehensive evaluation system. In-training assessment has the potential to measure the essential components of continuous professional performance, but difficulties encountered in assessment practice have raised serious questions regarding its utility.

The strength of existing models of in-training evaluation is that they are inexpensive and easy to use, although they do require some developmental resources and some time for rater training and implementation (Norcini & Burch, 2007). They provide an opportunity to evaluate continuous practice performance based on the assumption that the skills demonstrated are a realistic representation of the skills required for clinical practice. The areas of weakness relate to the psychometric properties associated with the tools, namely their reliability and validity. Moreover, unsystematic observation and documentation of learner performance, lack of rater consistency, the homogeneity of the population of interest, small sample sizes, and factors related to occasions for testing and timeliness of evaluations, all contribute to low reliability of scoring (Regehr et al., 2011). Also, content specificity as well as lack of discrimination between items suggest that the validity of the interpretations based on in-training scores is questionable (Eva, 2003). The infrastructure within which the ongoing assessment of students’ clinical competence operates may also be considered a weakness: time constraints that deter faculty from engaging in interaction and observation of students, and clerkships without clear objectives to indicate core-mastery-level material.

Given the challenges associated with current in-training evaluation practices, the meaningfulness of in-training methods as summative or formative assessment tools should be reconsidered. Although in-training evaluation comes closest to measuring true performance, current methods are unfortunately neither reliable
nor valid, and consequently unacceptable as a basis for trustworthy and defensible decisions about learners.

This thesis comprises several studies aimed at gaining insight into the use of the mini-CEX as a method of workplace-based assessment and the ways of improving its effectiveness for learning and assessment. The studies addressed the following research questions:

1. How reliable is the mini-CEX?
2. How valid is the mini-CEX?
3. How feasible is the mini-CEX?
4. What is the educational impact of the mini-CEX?

Chapter 2 addresses the validity, reliability, feasibility and satisfaction rates of this assessment tool in a naturalistic setting. In chapter 3 we assess the learning impact of the mini-CEX in a cohort of cardiology residents. In chapter 4, a study examining the teachers’ experiences on the role and functioning of the mini-CEX is described. In chapter 5 we made reliability estimations in a laboratory setup. In chapter 6 the learning impact of the mini-CEX in a cohort of medical students is examined and chapter 7 presents the relationship between judgment bias and types of scales. Since this dissertation is based on (published) articles, some repetition of information is inevitable.
References


Chapter 2
Validity, reliability, feasibility and satisfaction of the Mini-Clinical Evaluation Exercise (Mini-CEX) for cardiology residency training

A. Alves de Lima, C. Barrero, S. Baratta, Y. Castillo Costa, G. Bortman, J. Carabajales, D. Conde, M. Degrange, CPM van der Vleuten

*Medical Teacher* 2007; 29:785-790
Summary

The purpose of the study was to determine the validity, reliability, feasibility and satisfaction of the Mini-CEX. From May 2003 to December 2004, 108 residents from 17 cardiology residency programs in Buenos Aires were monitored by the educational board of the Argentine Society of Cardiology. Validity was evaluated by the instrument’s capability to discriminate between pre-existing levels of clinical seniority. For reliability, generalisability theory was used. Feasibility was defined by a minimum number of completed observations: 50% of the residents obtaining at least 4 Mini-CEX’s. Satisfaction was evaluated through a 1 to 9 rating scale from the evaluators, and residents’ perspectives. The total number of encounters was 253. Regarding validity, Mini-CEX was able to discriminate significantly between residents of different seniority. Reliability analysis indicated that a minimum of 10 evaluations are necessary to produce a minimally reliable inference, but more are preferable. Feasibility was poor: 15% of the residents were evaluated 4 or more times during the study period. High satisfaction ratings from evaluators’ and residents’ were achieved. Mini-CEX discriminates between pre-existing levels of seniority, requires considerable sampling to achieve sufficient reliability, and was not feasible within the current circumstances, but it was considered a valuable assessment tool as indicated by the evaluators’ and residents’ satisfaction ratings.
Introduction

The Mini-CEX has been designed to incorporate the skills that residents require in both actual patient encounters and in the educational interactions that they routinely encounter with attending physicians during teaching rounds (Holmboe, Huot, Chung, Norcini, & Hawkins, 2003; Norcini et al., 1995; Norcini, Blank, Arnold, & Kimball, 1997; Norcini et al., 2003). A single faculty member observes and evaluates a resident while that resident conducts a focused history and physical examination on an inpatient or outpatient, or in the emergency department setting. After asking the resident for a diagnostic and treatment plan, the faculty member completes a short evaluation form and gives the resident feedback (Holmboe, Yepes, Williams, & Huot, 2004). It is a performance-based evaluation method that is used to assess selected clinical competencies (e.g. patient charts and physical examination, also communication and interpersonal skills) in the medical training context. It is a performance-based assessment tool that intended to evaluate candidates at the “Does” level, that is, in real life settings and not in simulated situations (Miller, 1990). As the interaction is relatively brief and occurs as a natural part of the process in the training environment, each individual can be evaluated on several occasions and by various faculty members.

Assessment constitutes the most vital factor influencing student learning behavior (Newble, Hejka, & Whelan, 1990; Newble & Jaeger, 1983; van der Vleuten, 1996). When students see that the recall of factual information is a predominant requirement in the examination system, they tend to adopt a rote-learning or surface approach; however if examiners wish to assess students at the “Does” level, they must evaluate the student’s habitual performance in daily practice (van der Vleuten, 2000).

The major purpose of this study is to document if Mini-CEX applied in a broad range of clinical settings and in a big number of cardiology residency programs leads to achieving adequate levels of validity, reliability, feasibility and satisfaction rates from residents and teachers.

Methods

For each Mini-CEX, a single faculty member observed and evaluated the resident while the latter conducts a history and physical examination on an ‘in’ or outpatient or on a patient in the emergency department. After asking the resident for a diagnosis and treatment plan, the faculty member completes a short evaluation form and gives direct feedback. All formal mini-CEX evaluation data was collected on a one-page form that was the same in all of the different study sites where the study
was carried out. The form was previously translated and transculturally adapted into Spanish (Appendix 1).

Research subjects were cardiology residents from 17 cardiology training programs from Buenos Aires, Argentina. All the programs are affiliated to the University of Buenos Aires and consist of a four-year training period. The total number of residents of the entire program was 118. All the program directors were invited to participate. It was a completely new assessment strategy for all of them. Participation was voluntary, no incentives were provided. Written instructions about the application of the format were distributed and required at least 4 encounters during the 19 months study period. There were no sanctions for failing to participate, but all of them accepted to participate. The assessment was used as a maximum-performance but formative evaluation. Results were not used in evaluating residents for promotion.

Statistical analysis

Validity was evaluated by the ability of Mini-CEX to discriminate between pre-existing levels of expertise. In this case it was expected that significant mean resident performance differences are to be found between different years of training. The descriptive data were expressed as its means and standard deviations. For testing significance across expertise groups the non-parametric Mann Whitney test was used. A value of p < 0.05 was considered statistically significant.

To evaluate reliability, generalizability theory was used (Brennan, 2001). An ANOVA was carried out by identifying Year-of-training with (Y), Residents-within-Year with (P:Y) and Evaluations-within-Residents-within-Year with (E:P:Y), and variance components were estimated using the URGenova program. Since there might be significant growth in mean ratings throughout the years, variance associated with year was estimated separately (Y) to arrive at a more unbiased estimate of the variance of trainees. Separate evaluations of a single trainee could either be done by the same examiner or by a different one. This might have led to an underestimate of the variance across evaluations (intra-rater variability is probably smaller than inter-rater variability). Two indices of reliability were estimated using the variance components: Dependability Coefficients (D) and Standard Errors of Measurement (SEM), both as a function of the number of evaluations. The D-coefficient can be interpreted as a reliability coefficient, i.e. the expected correlation between other random evaluations of the number indicated using other evaluators and patients at random. The SEM is an estimate of the standard error and can be used to estimate confidence intervals around the score of an individual resident on the original scoring scale. For a 95% confidence interval the SEM is multiplied by 1.96 (z-score under which 95% of normal distribution lies). The SEM should be
below 0.26 (0.5/1.96) in order to produce a reliable inference on the scoring scale of at least one unit.

Feasibility was defined, according to the American Board of Internal Medicine's guidelines for Mini-CEX's implementations, on average a minimum of four Mini-CEX per resident (American Board of Internal Medicine, 2005)

Satisfaction was evaluated through the examination of the Mini-CEX from the perspective of evaluators, emphasizing on the ratings of the residents and on their satisfaction of the format. Ratings were carried out on a 9-point rating scale.

<table>
<thead>
<tr>
<th>1. Medical Interview skills (not observed)</th>
<th>2. Physical examination skills (not observed)</th>
<th>3. Reliable documentation/presentation (not observed)</th>
<th>4. Clinical judgment (not observed)</th>
<th>5. Counseling skills (not observed)</th>
<th>6. Objective / evidence (not observed)</th>
<th>7. Overall clinical competencies (not observed)</th>
<th>Mini-CEX score</th>
<th>Communication</th>
<th>Providing information</th>
<th>Evaluator satisfaction with mini-CEX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Number of times</td>
<td>None</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>None</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>None</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>None</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>None</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>None</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>None</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>None</td>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>

Appendix 1. The mini-CEX form
Results

From May 2003 to December 2004, 253 Mini-CEX encounters were carried out. 108 residents and 53 evaluators from 17 cardiology residency programs participated in the study. Each resident had gone through 1 to 7 evaluations (mean: 2.34): 13.7% of the residents were in their first year, 34.8% in the second, 41.2% in the third and 10.3% in their fourth year of residency. Each evaluator conducted between 1 and 21 (mean 4.77) evaluations. The total numbers of encounters were 253 and constituted the basis of the analysis. Of the 253 encounters, 52% of the encounters occurred in the coronary care unit setting, 30% were carried out in step down care unit, 6% in the emergency room, 6% in the ambulatory care unit and 6% in the cardiovascular intensive care unit. The overall competence ratings were similar in all settings. Forty-one percent of the encounters represented the first visit of a patient to a particular resident, and 59% were return visits to that resident. The mean total Mini-CEX time was 42.77 minutes (sd 19.97). If we divide the Mini-CEXs total time between assessment and feedback period, each individual assessment period takes 25.80 (sd 11.95, range 5-65) minutes and feedback period 17.31 (sd 11.28, range 5-65). The patient’s problems or diagnoses were specified by the evaluator and covered a broad range of problems in cardiology such as AMI, cardiac failure, unstable angina, atrial fibrillation, valvular disease and post-CABG. The mean ratings given by the 53 evaluators are reported in Table 1.

Validity Analysis

Validity was evaluated by examining if the instrument was capable of discriminating between pre-existing levels of clinical seniority. Mini-CEX discriminates between pre-existing levels of global competency between residents; first year residents 7.19 (sd 0.74), second 7.51 (sd 0.82), third 7.76 (sd 0.86) and fourth year residents 8.16 (sd 0.91), this difference reaches statistical significance. p= 0.0008. (Table 1)

Table 1. Mean Ratings given by all evaluators by year of training

<table>
<thead>
<tr>
<th>Domain</th>
<th>1 year</th>
<th>2 year</th>
<th>3 year</th>
<th>4 year</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication</td>
<td>7.16 (± 0.64)</td>
<td>7.57 (± 0.81)</td>
<td>7.57 (±0.92)</td>
<td>8.00 (±0.88)</td>
<td>0.002</td>
</tr>
<tr>
<td>Physical exam</td>
<td>7.12 (± 0.84)</td>
<td>7.48 (± 0.93)</td>
<td>7.59 (±0.96)</td>
<td>8.16 (± 0.91)</td>
<td>0.0006</td>
</tr>
<tr>
<td>Professionalism</td>
<td>7.64 (± 0.75)</td>
<td>7.82 (± 0.84)</td>
<td>7.83 (±1.00)</td>
<td>8.20 (± 0.93)</td>
<td>0.079</td>
</tr>
<tr>
<td>Clinical judgement</td>
<td>7.43 (± 0.71)</td>
<td>7.56 (± 0.86)</td>
<td>7.88 (±0.90)</td>
<td>8.20 (± 0.93)</td>
<td>0.0004</td>
</tr>
<tr>
<td>Counselling</td>
<td>7.43 (± 0.77)</td>
<td>7.44 (± 0.84)</td>
<td>7.59 (±1.11)</td>
<td>8.12 (± 0.90)</td>
<td>0.01</td>
</tr>
<tr>
<td>Organisation</td>
<td>7.32 (± 0.94)</td>
<td>7.54 (± 0.88)</td>
<td>7.68 (±1.01)</td>
<td>8.12 (± 0.90)</td>
<td>0.008</td>
</tr>
<tr>
<td>Global Competency</td>
<td>7.19 (± 0.74)</td>
<td>7.51 (± 0.82)</td>
<td>7.76 (±0.86)</td>
<td>8.16 (± 0.91)</td>
<td>0.0008</td>
</tr>
</tbody>
</table>
Reliability Analysis

The generalizability study theory yielded variance components for Y, P:Y and E:P:Y which were respectively 0.1643 (19.88% of total variance), 0.0482 (5.83%) and 0.614 (74.29%). Using the SEM benchmark of 0.26, a minimum of 10 evaluations were necessary to produce a minimal reliable inference (Table 2). This corresponds to a D-coefficient of 0.44.

Table 2. G-coefficients and SEM's are reported as a function of the sample size of evaluations.

<table>
<thead>
<tr>
<th>Number of Evaluations</th>
<th>G</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.07</td>
<td>0.78</td>
</tr>
<tr>
<td>2</td>
<td>0.14</td>
<td>0.55</td>
</tr>
<tr>
<td>5</td>
<td>0.28</td>
<td>0.35</td>
</tr>
<tr>
<td>10</td>
<td>0.44</td>
<td>0.25</td>
</tr>
<tr>
<td>15</td>
<td>0.54</td>
<td>0.20</td>
</tr>
<tr>
<td>30</td>
<td>0.70</td>
<td>0.14</td>
</tr>
<tr>
<td>50</td>
<td>0.80</td>
<td>0.11</td>
</tr>
</tbody>
</table>

G: generalizability-coefficient, SEM: standard error of measurement

Feasibility Analysis

For feasibility analysis the data showed that only 14.81 % of all the cohort was evaluated 4 or more times during the study period.

Satisfaction Analysis

The residents (108) were generally satisfied with the mini-CEX format; their ratings ranged from 5 to 9 (mean 8.08 ± 0.83). Satisfaction rate for first year residents was of 8.1, second year residents 7.8, third year residents 8.1 and fourth year residents 8.5. The evaluators were also satisfied with the mini-CEX format; their ratings ranged from 6 to 9 (mean 8.06 ± 0.74).

Discussion

The purpose of the study was to report logistic and psychometric data for the Mini-CEX format. Regarding validity, Mini-CEX clearly was able to discriminate between pre-existing levels of global competency between residents, has insufficient reproducibility when only a few evaluations are sampled and had high satisfaction rates.
according to both evaluators and resident’s satisfaction rates. Regarding feasibility, it appeared not possible to achieve the number of encounters required.

Some issues may explain this difficulty. It was a new assessment tool never applied earlier in this environment. We only developed written instructions, and perhaps in vivo faculties training programs for Mini-CEX would be preferable. Regarding reliability, one-fifth of the variance can be attributed to growth in competence throughout the years. Provided that the Mini-CEX should offer information on growth towards a final level of competence and should be able to discriminate throughout years of training. Only approximately 6% of total variance is related to resident (or person) variance. Since the instrument is designed to discriminate among residents, this constitutes desirable (true score or universe score) variance and the larger it is the better. As usual in many competence and performance assessments, however, this component is relatively small (van der Vleuten, 2005).

About three-quarters of variance is associated with differences between examiners/evaluation occasions and residual error. Using the SEM benchmark of 0.26, a minimum of 10 evaluations are necessary to produce a minimal reliable inference. The D coefficient for 10 observations was, however, rather low (0.44). Reliability coefficients of 0.80 or higher are generally accepted as a threshold for high-stakes judgements, such as the registration of a doctor for licensure (Crossley, Davies, Humphris, & Jolly, 2002). In our dataset this was achieved with 50 observations.

The required number of encounters as derived from this study is more demanding than reported by other studies in the literature. Norcini concluded that ten or more encounters produced relatively tight confidential intervals and that an increase in the number of encounters beyond that produced only small gains in consistency (Norcini et al., 2003). Carline concluded that 7 independent ratings would be necessary to judge overall clinical performance (Carline, Paauw, Thiede, & Ramsey, 1992). Similar results have been reported by Kreiter (more than 8 ratings needed), Kwolek (7-8 ratings), Ramsey 11 ratings, Violato 10 ratings and Kroboth 6-10 ratings (Burning, Cation, Markert, & Pangaro, 2002; Kreiter, Ferguson, Lee, Brennan, & Densen, 1998; Kroboth et al., 1992; Kwolek, Witzke, Blue, Schwartz, & Sloan, 1997; Violato, Marini, Toews, Lockyer, & Fidler, 1997; Wenrich, Carline, Giles, & Ramsey, 1993). All reports agree that somehow between 7 and 11 ratings are necessary to achieve a generalizable global estimate of competence when ratings are based on a non-systematic sample of observations (Williams, Klamen, & McGaghie, 2003). Although there has been only limited research on the component skills included under the broad category of clinical competence, it is reasonable to expect that these abilities will develop at different rates and may differ in their stability in different situations. Results of performed work suggest that different numbers of observations will be required to establish stable estimates of clinical competence in various clinical competence areas (Williams et al, 2003).
We have not analyzed inter-rater reliability since we took a time sampling perspective in which each observation is but one observation in a longer time-framework. We analyzed reliability across these observations. Rater variance as well performance variability variance across observations will be part of our reliability estimation (although we can’t partition out the two sources of variance). If only one observation is used for a Mini-CEX assessment it would be important to judge the inter-rater reliability. Our data, unfortunately, do not provide this information. Consistent with previous works, examiners were satisfied with the format (Williams et al., 2003).

**Limitations of the study**

The number of residents participating in our study was relatively low and this group may perhaps not be fully representative of broader populations.

We are aware that improvement of performance across years of training is a weak form of construct validity. However, it is a fundamental one: absence of performance differences across different expertise groups would be detrimental to construct validity of the instrument used. This not being the case, we conclude to a first indication of validity. Further studies into construct validity should be the next step. Studies looking at the incremental validity over existing more standardized performance instruments would provide compelling construct validity evidence.

The assessment used was a maximum-performance but formative evaluation. If we take into account that this assessment does not assign grades or certifications, this could seriously have affected the perception of the fellows and influence their satisfaction rates.

The reliability analysis carried out here used the usual assumption of local independence between the repeated measurement moments (one measurement is not influenced by another measurement). This assumption is clearly violated in Mini-CEX studies including ours. Every evaluation is actually meant to provide feedback and to change the performance of the person being assessed. The Mini-CEX evaluations are therefore not independent of each other. This is a general problem in the literature of performance measure which are dispersed in time and used for formative purposes (Alves de Lima et al., 2005).

**Conclusion**

The direct observation of residents behaviour is essential to assess clinical skills (Holmboe, 2004; Holmboe et al., 2004; Schuwirth & van der Vleuten, 2006). For decades, clinical supervisors have taken at face value the veracity of the history and
physical examination presented on inpatient and outpatient rounds without ever directly observing how the trainee actually performed them (Holmboe, 2004).

Medical educators’ major challenge lying ahead is to find the way to ensure that they themselves have not only strong clinical skills, but also the necessary skills to effectively observe, evaluate, and provide constructive feedback to trainees regarding clinical skills. In this direction Mini-CEX ensures direct observation and feedback of residents from different faculties in a broad range of patients’ problems in various settings. Furthermore, as this demonstrated feasibility is an issue. Application strategies should be reinforced. We don’t think that Mini-CEX requires any modification in itself but it should never be used as a unique assessment tool. Direct observation of trainees in clinical setting can be connected to other exercises that trainees may perform after their encounters with patients, such as oral case presentation, written exercises that assess the clinical reasoning, and literature searches. In addition review videos of encounters with patients offer a powerful means of evaluating and providing feedback on trainees’ skills in clinical interaction (Epstein, 2007). It can only be approve if this evaluation method becomes part of the clinical routine of the clinician and resident.
References


Chapter 3

A qualitative study of the impact on learning of the mini clinical evaluation exercise in postgraduate training

A. Alves de Lima, R. Henquin, J. Thierer, J. Paulin, S. Lamari, F. Belcastro, CPM van der Vleuten

Medical Teacher 2007; 29:785-790
Summary

The study was designed to illustrate how residents perceive the Mini Clinical Examination Exercise as an assessment tool and its influence on their approach to learning and studying. A phenomenographic approach was applied. All 16 residents from a cardiology training program in Buenos Aires were included. Results show that in all cases residents demonstrate an intrinsic interest in the subject matter. They show self-regulating strategies when required to select, relate and make critical appraisals of their own. They consistently demonstrate an aim to build a relationship between individual experience and their chosen topic. The residents feel comfortable because it melds with their routine. Residents find the Mini Clinical Examination Exercise to be a useful assessment tool with a favorable influence towards a constructive approach to study and learning.
Introduction

The Mini Clinical Evaluation Exercise (Mini-CEX) aims to assess students at the top of Miller’s pyramid (Miller, 1990; Norcini et al., 1995, 1997; Norcini et al., 2003). Assessment literature in general has shown positive results at almost every level in the ‘climbing’ of Miller’s pyramid. The only exception to this ascent is at the top, which is particularly relevant for senior-year medical trainees in postgraduate situations. The Mini-CEX has been designed to incorporate both the skills that residents require in actual patient encounters and with the educational interactions that attending physicians routinely encounter with residents during teaching rounds. It is an evaluation method that promotes the assessment of clinical skills along with attitudes and behaviors that are essential in high-quality patient care. It is a performance-based evaluation method that is used to assess selected clinical competences (e.g. patient charts and physical examination, also communication and interpersonal skills) in the medical training context. A single faculty member observes and evaluates a resident while he/she conducts a thorough history and physical examination on an in- or outpatient or a patient in the emergency department. After asking the resident for a diagnosis and treatment plan, the faculty member completes a short evaluation form and gives direct feedback. As the interaction is relatively brief and occurs as a natural part of the process in the training environment, each individual can be evaluated on several occasions and by various faculty members. Ten or more encounters are required in order to reach a reproducibility of 0.80. According to Norcini, data collected on confidence intervals is important since it makes it clear that, even with relatively few encounters, useful information can be gathered (Norcini et al., 1995). Confidence intervals provide additional information that permits the test length to be shortened and tailored to specific situations (Norcini et al., 1995). Performance-based evaluations allow for the medical trainee to perform in a real-life situation and according to Miller’s pyramid it is at the ‘shows-how’ level in which trainees reflect their ability to act appropriately in a practical situation (Miller, 1990).

Undoubtedly, the assessment system constitutes the most vital factor influencing student learning behavior (Newble et al., 1990; Newble & Jaeger, 1983). When students see that the recall of factual information is a predominant requirement in the examination system, they tend to adopt a rote-learning or surface approach; however, if examiners wish to assess students at the highest level of Miller’s pyramid, they must evaluate the student’s habitual performance in daily practice (van der Vleuten, 2000). We wanted to study whether a measure from the highest level of the pyramid can influence the way that residents approach their studies. We chose the Mini-CEX because as a performance-based tool it encourages the resident to resolve ill-structured problems, driving learning to a deeper level. We expect the
Mini-CEX format assessment to have a more favorable influence on those with a tendency toward a deeper approach to learning.

Methods

The study took place at the Cardiovascular Institute of Buenos Aires (ICBA), a 55-bed cardiovascular training hospital located in the federal district of Buenos Aires province, Argentina. The institution itself, together with the cardiology residency program, is affiliated with the University of Buenos Aires (UBA).

Subjects

Research subjects were cardiology residents from the training program. The program consists of a four-year training period with 16 residents evenly stationed (four per year). All 16 were invited to participate. Participation was voluntary and every resident accepted.

The phenomenographic approach

As the aim of the study was to focus on how residents perceive an assessment tool and its influence on their approach to learning, an inductive method was used. With this in mind, the analysis was carried out in accordance with the phenomenographic approach, which involves repeated readings of the transcribed interviews in order to gain a thorough understanding of each interview. The analysis was conducted through the relation and comparison of each sector of each interview with the relevant and corresponding sectors of other interviews (Marton, 1979). Similarities and dissimilarities of the components of the interview that emerged during the analysis were categorized according to description (Marton, 1986; Marton & Svensson, 1979; Elder, 1995; Stacy, 2000).

Implementation procedure

So as not to interfere with daily activities, the following procedure was developed: the exam (Mini-CEX format) was scheduled for 28 May. Each participant received a memorandum with the instructions 45 days before the exam. Each resident participated in only one encounter. The interview plan consisted of one interview per participant during the 72 hours following the exam (Figure 1). The Mini-CEX provides a 15 to 20 minute snapshot of a resident/patient interaction which focuses on the core skills that residents demonstrate in patient encounters (Norcini et al.,
Four examiners participated in the study. Each examiner was randomly assigned to four residents. All four examiners were attending physicians from the institution. All of them had received training on the application of this format. Only one examiner was aware of the aims of the study. The assessment used was a maximum-performance but formative evaluation.

Figure 1. Study design

Descriptors of competencies demonstrated during the Mini-CEX

In relation to the skills demonstrated during Mini-CEX we would mention (figure 2):

- Medical Interviewing Skills: facilitates patient rapport through effective use of questions to obtain accurate and adequate information necessary; responds appropriately to non-verbal cues.
- Physical Examination Skills: Follows logical sequence, efficiency; balances screening/diagnostic steps relevant to problem; thoroughly informs patient and is sensitive to patient’s comfort and modesty.
- Humanistic Qualities/ Professionalism: Shows respect, compassion, empathy, establishes trust; attends to patient's necessities: comfort, modesty, confidentiality, information.
- Clinical Judgment: Selectively orders/ performs appropriate diagnostic studies, considers risks and benefits.
- Counseling Skills: Explains rationale for test/ treatment, obtains patient consent, educates/counsels on management.
- Organization/ Efficiency Skills: Prioritizes; is timely; is succinct.
- Overall Clinical Competence Skills: Demonstrates judgment, synthesis, care, effectiveness, efficiency.
Open and semi-structured interviews constitute the basis for gathering data within the interview plan. The reason for choosing the interview as a data collection method is that it was deemed valuable to give the residents the opportunity to describe how the concept of their preparation and approach to learning prior to an examination was perceived.

An interview outline introduced a number of phenomena by considering the following fields or domains (figure 3):
• Preparation strategies
• Regulation strategies
• Affective strategies
• Appraisal of the exam method

1. Cognitive learning (preparation) strategies
   1.1. How did you prepare for this exam?
   1.2. How did you begin your preparation?
   1.3. Can you describe for me, in as much detail as possible
   1.4. How did you organised your activities?
   1.5. What kind of activities did you organised?
   1.6. Did you feel that some tasks were too difficult to tackle?
   1.7. If, yes, what was your strategy?
   1.8. Did you read the instructions given before the exam?
   1.9. Did you take into account the assessment format at the moment that you are were preparing for the exam. In what way, how?

2. Regulation of preparation (metacognitive regulation strategies)
   2.1. How did you know if you had achieved an adequate level of knowledge in order to succeed in the exam?
   2.2. What was your strategy?
   2.3. How did you organise your priorities? Did you have a basis for this?
   2.4. Did you take into account things that were taught during the meetings?
   2.5. Did you prioritise what the teacher considered a priority?
   2.6. For this format, did you consider it important to have previous experience in this format in order to succeed?

3. Affective strategies (like motivating, regulating own emotions, etc.)
   3.1. How did you feel prior to the exam?
   3.2. Do you think that the assessment format could interfere with your anxiety level?
   3.3. Do you think that anxiety could interfere with your results?
   3.4. What is the relationship between the themes and your practice?
   3.5. How can you explain this?

4. Appraisals of the exam format
   4.1. How do you feel about this format, did you like it?
   4.2. Do you feel that this is an adequate assessment technique?
   4.3. Did you feel comfortable?
   4.4. Can you name some favourable characteristics of this format, please?
   4.5. Can you name some unfavourable characteristics of this format?

Figure 3. Interview format
Data analysis

In the study, interviews were initially recorded and transcribed literally by an independent typewriter leading up to analysis. The researcher then read each interview a number of times, gaining an overall impression. Subsequently, interviews were processed according to the concepts taken from residents’ statements. The core of analysis entailed the comparison of different statements, revealing both similarities and differences. In order to obtain an overview of how these similarities and differences could be connected, they were grouped into patterns. The final pattern to emerge resulted in categories that were obtained through the interaction between the segments as a whole and their parts (Elder & Miller, 1995). The researcher was involved as interviewer and in some cases as examiner. Practical circumstances avoided stronger methodological strategies.

Results

The results are presented in four categories within the interview

• Preparation strategies;
• Regulation strategies;
• Affective strategies;
• Appraisal of the exam methods;

The analysis of the transcripts leads to the identification of different categories in each section. As mentioned earlier in the methods section, emphasis is placed on the qualitative description of the categories (Dahlgren, Diwan, Tomsom, & Wahlstrom, 1992). Feedback on preparation, regulation, affective strategies and the concept of the format itself originate from the context of each individual's experience but are not attributable to one single resident. The results are holistic, describe common experiences and are a collective understanding of the qualitative and varying ways that residents experience assessment formats. Several times during the interviews, residents expressed more than one concept, but not all of them shared the same experience or understanding of the phenomenon.

Reflections on preparation strategies

The first question was related to preparation strategies for the Mini-CEX format as a general phenomenon. Participants were asked for a spontaneous reflection and their answers were grouped into two categories: (a) preparation in terms of understanding and construction of their knowledge and (b) preparation in terms of training skills with everyday practice.
(a) Preparation in terms of understanding and constructing their own knowledge.
This category describes the residents’ conception of their ability to relate elements of the subject matter to each other and to incorporate these elements as a whole when aiming to link the topics to their practice:

*I tried to construct an overall picture of the topics, based on the most frequent patients that are admitted in my sector...*

*I tried to compare the different points of view in relation with the topics of the exam. Although I took into account the information given in meetings and rounds, I tried to develop my own overall view...*

(b) Preparation in terms of training skills with everyday practice.
In this category residents describe that the format is related to their everyday practice.

*I read the instructions but I didn’t organise anything special. The performance fashion of the format is my daily practice...*

**Reflections on regulation strategies**

The category describes residents’ concepts relating to how their realization of whether they have achieved the adequate knowledge to succeed in the Mini-CEX format. The participants’ answers could be grouped into one category: regulation as the ability to explain and perform the skills learned.

(a) Regulation in terms of the ability to explain and perform the skills learned.
This category describes the residents’ conception about regulation as a personal understanding process:

*In this case I have to think, to explain orally and also to show my skills, and I think that I have achieved an adequate level when I am able to perform all of the tasks... No, I don’t need to train on the format itself, because I am working on “the format” throughout my entire journey...*

**Reflection on affective strategies**

All residents agree that anxiety is the main problem that this format can produce. Anxiety can either result in a hindering feeling or become a motivational effect. The answers given fall into the category of affective strategies, represented as a feeling of anxiety.

(a) Affective strategies as feeling of anxiety.
This category describes the concept of external influence on the format. It indicates a potential interference in the participants’ performance during the exam. In this
format the need to interact with the examiner can promote a negative effect. All residents agree that the interaction with the examiner enhances anxiety. *During oral exams you have to discuss your ideas with another person. There are factors that can interfere such as: personality or halo effects... It is my environment. It is the way that I have always worked.*

**Reflections on the format itself**

The answers provided by the residents fall into two categories: (a) a gaining in understanding of one's level of expertise and (b) an educational experience.

(a) A gaining in understanding of the residents level of expertise

This category describes the conception that the format assesses residents in a 'real world' environment. Residents experience what the format takes into account from their environment.

*It evaluates our everyday activity... I think that is the best way evaluation method...*

(b) An educational experience

This category describes the moment in which the residents experience just how useful this assessment tool is in helping to enhance their knowledge. This happens when residents are able to share and interpret information during the assessment process. The feedback also gives them a clear performance update:

*It's an educational or learning experience... When I say something wrong, I have the possibility to explain the reasons for my arguments in relation to the topic...*

In summary, most of the conceptions that residents seem to have regarding preparation for the Mini-CEX are associated to deeper learning characteristics.

**Discussion**

The purpose of the present study was to evaluate cardiology residents’ approaches to learning and studying when exposed to a performance-assessment instrument during their training program in a teaching hospital. Here follows a summary of major conclusions we have reached regarding the four categories mentioned above:

(a) In relation to preparation strategies, which refer to the thinking process that a fellow needs in order to assimilate learning previous to the exam, students have expressed that: they always try to understand the meaning of the subject matter. They look for central concepts and try to prove or reject them in relation to their opinions and experiences. "Preparation” therefore, is seen as an under-
standing and construction-based process of their knowledge in order to succeed. This precisely, is a characteristic of a deep learning approach.

(b) In relation to regulation strategies, residents adapt their approach to study according to their personal interest, prior knowledge and requirements, which are, in their view, needed for the fulfillment of the task. This type of regulation is internally oriented or self-regulated and is also characteristic of a deep-learning approach.

(c) In relation to effective learning activities, residents consistently demonstrate an intrinsic interest in the topics studied for exam. They attempt to construct a relationship between personal experience and the topics they study to achieve set objectives. This is another characteristic of a deep-learning approach.

(d) The fourth category relates to the reflection of the format itself. Each individual feels quite comfortable with the Mini-CEX as it allows for the evaluation of hands-on practice. Students agree that it constitutes a valuable assessment strategy and is a motivating educational experience.

Results show that the residents, in response to: preparation strategies, regulation strategies and effective learning activities, consistently demonstrate intrinsic interest in the subject matter regardless of the assessment method. They aim to understand the meaning of the subjects. They often describe their goal as being the interrelation of all parts of the subject matter to construct a broader picture of the problem. In every case, residents reveal self-regulating strategies at the moment of selecting, relating and making critical appraisals of their own. The reflection on the format itself shows that they feel very comfortable with the performance-based format because it is intertwined in their routine.

Conclusion

The assessment’s objectives should clearly match educational objectives. When they do not, the assessment objectives will prevail. The implication for practice is to be constantly aware of the educational effects of assessment and to try to use the driving forces of assessment in order to achieve desirable educational effects. Examination motivates students’ learning. This law illustrates one of the strongest relationships in education. If examiners wish to assess students at the highest level of Miller’s pyramid they must not give them tests of memory reproduction. They should give tests to evaluate the student’s habitual performance in every day practice. In this vein, the Mini-Clinical Examination Exercise is a performance-based format, designed to introduce the direct observation of the trainee’s clinical competence and to assess the integration of clinical skills. It is a valuable teaching and feedback tool that orients residents to a more fully integrated approach to learning.
It is only through research into residents’ learning styles and approach to learning that the evidential basis for educational change can be provided. Before this, any justification lies on intuitive grounds. Moreover, it is essential to consider how one conceives their leaning and studying since this characterizes, to a great extent, the teaching provided. Continued research would therefore be of interest regarding this topic. The implication from a practical point of view is to be constantly aware of the educational effects of assessment and to try to use the driving forces of assessment to achieve desirable educational effects (van der Vleuten, 1996).

**Limitations of the study**

Although our outcomes are in accordance with our expectation, it is also possible to find other factors that may have had an influence.

The assessment used was a maximum-performance but formative evaluation. Maximum performance assessments have been labeled as those procedures used to determine a person’s ability. They are concerned with how well individuals perform when they are motivated to obtain the highest possible score. Assessment results indicate what individuals can do when they make their best effort and formative assessments are used to monitor learning progress during instruction. Their final purpose is to provide continuous feedback to both students and teachers concerning successful learning and to identify the specific learning errors that would need correction. However, if we take into account that this assessment does not assign grades or certifications, this could seriously have affected the perception of the fellows and influence their approach to learning a studying.

All of the participants were cardiology residents. They joined our program by their own decision. Furthermore, keeping up to date as well as maintaining competence constitutes dependent voluntary self-motivated activities which are likely to be undertaken most effectively by those practicing a deep learning approach. The fellow training programs themselves may constitute an effective educational technique in encouraging residents to shift from a superficial learning approach to a deep learning approach (Newble et al., 1990). In unpublished data, we have recorded the learning style assessment of 107 cardiology residents of the Argentine Society of Cardiology through the application of the inventory learning style developed by Vermunt and found that they showed a significantly higher score on a deep approach to learning in all of the four categories evaluated (Vermunt, 1996).

Physicians in this study were recruited from the upper academic quarters of medical schools graduates. These successful students might be those with more desirable approaches to learning (Newble et al., 1990).
References


Chapter 4

Teachers’ experiences of the role and function of the mini clinical evaluation exercise in post-graduate training

A. Alves de Lima, D. Conde, L. Aldunate, CPM van der Vleuten

*International Journal of Medical Education 2010; 1:68-73*
Summary

The mini clinical evaluation exercise is a performance-based evaluation method that is used to assess, through feedback on observed performance, residents’ clinical competencies, such as medical interviewing skills, physical examination, and clinical judgment, within a medical training context or single patient encounter. The objectives of the study was to investigate teachers’ experiences and views regarding the mini clinical evaluation exercise as an assessment tool and to evaluate its feasibility as well as its influence on teachers’ pass/fail decisions and feedback delivery. Seventeen teachers who had all used the mini clinical evaluation exercise in assessing residents at least twice during the study period were interviewed. Transcripts of the interviews were analyzed qualitatively using a phenomenographic approach. All teachers considered the mini clinical evaluation exercise as a useful assessment tool that promotes direct observation and constructive feedback. The format was considered feasible because of its easy adaptability to daily practice. Uncertainty as to what should specifically be observed during encounters, interpersonal relationships, and preconceived notions of resident performance were given as reasons for teachers’ difficulties in defining a pass or fail score. Teachers generally tended to be averse to failing residents. This study shows that teachers perceive the mini clinical evaluation exercise as a feasible and useful formative assessment tool. Contextual factors such as interpersonal relationships, preconceived performance notions, and lack of specific guidelines and performance standards appear to explain why teachers tend to be reluctant to fail poorly performing residents.
Introduction

The mini clinical evaluation exercise (mini-CEX) is designed around both the skills most commonly performed by residents in actual patient encounters and the educational interactions that attending physicians routinely have with residents during teaching rounds. It is a performance-based evaluation method that can be used to assess selected clinical competencies (e.g., physical examination, communication and interpersonal skills, professionalism) in the clinical training context (Norcini et al., 2003).

In educational practice, decisions are rarely based on research and, especially with regard to assessment, teachers, students, and institutions tend to have strong opinions that are largely based on sentiments and tradition. Teachers are usually unaware of educational research or do not consider it important (Nelson, Clayton, & Moreno, 1990). The extent to which an assessment procedure is accepted by those involved in its execution is a crucial element in the introduction of a new assessment method.

Several studies have focused on teachers’ and students’ experiences with the mini-CEX. Norcini and colleagues showed that examiners as a group were very satisfied with this method (Norcini et al., 2003). Torre and colleagues showed that the mini-CEX was highly rated by students and evaluators as a valuable tool to document direct supervision of clinical skills (Torre, Simpson, Elnicki, Sebastian, & Holmboe, 2007).

The extent to which an assessment procedure is accepted by those involved in its execution is a crucial element in the introduction of a new assessment method. Alves de Lima et al. (2007) also found that residents and evaluators were satisfied with the mini-CEX format (Alves de Lima, et al. 2007). Residents’ ratings ranged from 5 to 9 on a ten point scale (mean 8.08 ± 0.83) and evaluators’ ratings ranged from 6 to 9 (mean 8.06 ± 0.74). Assessment procedures that are not accepted by teachers or students are likely to be discarded eventually. Probably, the best course of action in achieving acceptance is to make strategic use of information about teachers’ and students’ beliefs in order to gain their commitment (van der Vleuten, 1996). This seems particularly applicable with work-based assessment where the value of the assessment appears to be determined by the users of the instruments rather than by the instruments themselves (van der Vleuten, Schuwirth, Scheele, Driessen, & Hodges, 2010). The purpose of this study was to investigate how teachers value the mini-CEX as an assessment tool, its feasibility, the influence of contextual factors on teachers’ pass or fail decisions, and feedback delivery. We conducted a qualitative study in which we interviewed teachers about their experiences with the mini-CEX. In analyzing the interview data we used a phenomenographic approach in order to gain a thorough understanding of teachers’ experiences. Unlike
many other assessments, where the teacher is largely a passive instrument, the mini-CEX requires teachers to actively construe their judgments. We need to understand this process as well as contextual factors that influence teachers’ judgments not only in terms of feedback but also in terms of summative decision making. We expect that such insights will promote the use of the mini-CEX and help to optimize the format (Govaerts, van der Vleuten, Schuwirth, &Muijtjens, 2007). The study was conducted at the Cardiovascular Institute of Buenos Aires (ICBA), a 55-bed cardiovascular training hospital located in the federal district of Buenos Aires, the capital of Argentina. Both the institution and the cardiology residency program are affiliated with the University of Buenos Aires (UBA).

**Methods**

**Participants**

The participants were cardiology teachers delivering the postgraduate cardiology training program. The following criterion was used to select candidates: faculty members who had used the mini-CEX to assess residents on at least two occasions between May 2005 and May 2006. Out of 24 teachers, 17 met this criterion and we invited these to participate in the study. All 17 agreed to participate. Participation was voluntary. Three, eight, and six teachers had used the mini-CEX twice, three to five times, and more than five times, respectively. Three had assessed the same resident more than once. In 2002 the mini-CEX was introduced in the cardiology residency program for formative assessment. Since 2004 it has been used for summative assessment and all the participants had used the mini-CEX during this period. In 2004 the program director met with each assessor to discuss general performance standards. The residency program of the Cardiovascular Institute of Buenos Aires lasts four years and the Institute accommodates four residents at the same time. Every year each resident takes part in five mini-CEX assessments. Residents who fail have one chance to re-do the mini-CEX, only this time not with their teacher but with the program director. For the mini-CEX, one faculty member observes and evaluates a resident who is taking a history and performing a physical examination on an inpatient, an out-patient or on a patient in the emergency department. After the encounter the resident presents a diagnosis and management plan to the faculty member, who then completes a short evaluation form and gives feedback. The mean time used for the mini-CEX was 42.77 minutes (SD=19.97). Residents’ performance is rated on the following competencies: medical interviewing skills, physical examination skills, humanistic qualities, clinical judgment, counseling skills, organization skills and efficiency, and the resident’s overall clinical competence. Ratings are given on a 9-point scale: 1, 2, and 3 indicate unsatisfactory
performance, 4 marginal performance, 5 and 6 satisfactory performance, and 7, 8 and 9 indicate superior performance. A rating of 1, 2 or 3 means a fail. In addition to the data on the resident’s performance, the teacher also records the site of the assessment (the inpatient service, the out-patient clinic, or the emergency department), the complexity of the case (low, moderate, high), and the patient’s gender, age, and main medical problems and diagnoses. This research protocol was approved by the institutional Review Board of the Cardiovascular Institute of Buenos Aires.

**The phenomenographic approach**

We used an inductive method to explore teachers’ experiences with the mini-CEX as an assessment tool, its feasibility, and its influence on their rating and feedback delivery strategies. Using a phenomenographic analytical approach we repeatedly read the transcriptions of the interviews in order to gain an in-depth understanding of each interview (Marton, 1986).

**Analysis**

We designed a procedure for the study aimed at minimizing interference with daily clinical routine. Thirty days before their interview participants received a written invitation. We conducted one open, semi-structured interview per participant and continued interviewing until we considered that saturation had been reached. We used interviews to collect data, because this method would allow each participant to elaborate on his/her perceptions of the mini-CEX. We used an interview guide (see Figure 1) to elicit the participants’ views regarding the mini-CEX with regard to its feasibility, teachers’ strategies to arrive at pass/fail decisions, teachers’ appraisal of the format, and feedback. One of the investigators conducted the interviews, which lasted, on average, 55 minutes (range 35-80 minutes).
1. **Implementation strategies**
   1.1. Describe your experience with respect to the implementation of the format, in detail.
   1.2. How did you organize the activity? (Schedule)
   1.3. Explain the reasons for difficulties (if any) arising from the implementation of the format.
   1.4. Did you feel that the task was very difficult to tackle?
   1.5. If so, what was your strategy?
   1.6. Did you read the instructions given before the exam?
   1.7. Did you follow the instructions? Reasons for doing / not doing so.

2. **Rating strategies**
   2.1. Do you think that the exam format keeps residents from failing?
   2.2. In case you failed a resident: were you sure about your decision?
   2.3. In what case did you give the resident the benefit of the doubt?
   2.4. Were you sure about it?
   2.5. How badly should a resident do to fail?
   2.6. Did this format influence your rating behavior?
   2.7. In what way?
   2.8. Do you think it is important for residents to have previous experience on this format in order to succeed?

3. **Appraisals on the exam format**
   3.1. What is your opinion about the format itself?
   3.2. Do you think it is a good assessment tool? Why?
   3.3. How did you feel about this format, did you like it?
   3.4. Did you feel comfortable?
   3.5. Did the format alter your daily practice?
   3.6. Do you think that this format influences students approach to learning and studying. In what way?
   3.7. Could you give some examples, please?
   3.8. Please, name some favorable characteristics of this format.
   3.9. Please, name some unfavorable characteristics of this format.

4. **The nature of the feedback**
   4.1. Could you please provide a detailed report on your feedback? (recommendations for improvement, learner reaction, self-assessment, action plan)
   4.2. How important do you consider this activity?
   4.3. How do you feel about it? Do you feel comfortable by carrying it out?
   4.4. Which are the main strengths of the feedback in this format?
   4.5. Which are the main weaknesses of the feedback in this format?

---

**Figure 1. Interview outline**

We audio recorded interviews after obtaining consent from each participant. We then transcribed the recorded interviews and two investigators independently coded the transcripts. In analyzing the transcripts, the investigators related and compared the sections of each interview with the corresponding sections of the other interviews (Marton & Svensson, 1979). The investigators categorized the similarities and dissimilarities between the interviews that were identified during the analysis (Elder & Miller, 1995; Stacy & Spencer, 2000).
Results

The results are presented for each of the four issues addressed during the interviews (see Appendix). The results are holistic. They represent common experiences recounted by teachers and reflect the different ways in which the teachers experienced the mini-CEX assessment format.

Feasibility issues

The participants’ views are grouped into one category: integration of the method in patient care and integration of teaching activities in daily practice.

(a) Integrating patient care and teaching activities in clinical practice
All the teachers worked in very busy clinics and struggled to combine clinical and administrative duties with their teaching responsibilities. All the teachers agreed that the mini-CEX was a very useful assessment instrument, and they were looking forward to its further implementation. For most of them, the main challenge was to incorporate direct observation of residents into practice routines, something they were definitely not used to doing. The majority of the teachers were used to assessment in which they listened to residents presenting their findings from and records of physical exams. The teachers said that observing residents during history and physical examination enabled them to detect important errors in residents’ performance. The teachers used two different major implementation strategies. Some teachers scheduled appointments for the mini-CEX in their daily schedules, while others preferred on the spot observation when deemed appropriate. Some teachers agreed upon a schedule with the resident and some even asked residents to remind them of the arranged date and time of the mini-CEX. The teachers suggested that a pocket-sized form would be of great help.

No, there weren’t any great difficulties in the implementation; the major difficulty is scheduling time…..to find some quiet time to devote to this type of protocol...

Pass/fail strategies

This item relates to teachers’ feelings when they had to decide whether a resident’s performance merited a pass or a fail. The participants’ answers could be grouped into two categories: feelings of discomfort or insecurity in making this decision and their perception that the format could induce avoidance of fail decisions.

(a) Feelings of discomfort or insecurity in making pass/fail decision
Most of the teachers were confident of their capability to determine whether a resident was performing poorly, but they had difficulty identifying the pass/fail thresh-
old. They did not know which type of information or which specific behaviors they should document to support a judgment of poor performance. As a consequence, they tended to avoid asking questions until the resident had reached a satisfactory level of performance during the assessment. They felt uncomfortable giving pass/fail judgments, and they all agreed that, in borderline cases, they preferred giving a resident the benefit of the doubt. This could lead to inflation of scores and false positive decisions.

Yes, it depends on where the pass/fail point is considered to lie...but I guess that it is much more flexible; as there are many items... someone may therefore fail one item and do very well on another one...

(b) Avoidance of fail decisions as a result of using the mini-CEX
Teachers reported difficulty establishing which level a resident was expected to have attained, due to heterogeneity among residents of the same year. Some residents do well on certain aspects, but perform poorly on other aspects. As a result, an overall competence judgment requires a delicate process of weighing different pieces of information and combining them into an aggregate decision. The teachers indicated that they perceived this to be a particularly delicate and difficult task.

In addition to the intrinsic challenges of the assessment task, the teachers mentioned the impact of personal relationships between teachers and residents. The teacher generally knows the resident and, consequently, has a pre-conceived notion of his or her competence. Also, the relationship of residents with the Institution plays an important role: the teachers felt that it was unacceptable to fail a resident whom one had personally selected for admission to the Institution's residency program.

I always know something about the resident's clinical competence, if the resident shows poor performance I start asking new questions because I am sure that he will do well...

Format issues
The teachers' views on this topic fell into two categories: assessment in an authentic environment and positive educational experiences.

(a) Assessment in an authentic environment
The majority of the teachers indicated that the format enabled them to assess all aspects of clinical practice. In contrast to other formats, such as MCQ or essay questions, the mini-CEX allows realistic cardiologic clinical scenarios to be addressed in the assessment. Accordingly, realism was described as an important aspect of the use of real patient cases in an authentic clinical environment. Also, teachers valued the opportunity to confront residents with cases of different complexity.
As opposed to multiple choice assessments where only theoretical knowledge counts, this type of assessment evaluates the doctor in action... (Interviewee T) 2.1)

(b) A positive educational experience
Many teachers valued their exposure to students’ work and the opportunity to review students’ clinical skills, such as physical examination and communication skills, which was afforded by the mini-CEX. They found it a useful teaching tool especially as their feedback and direct observation skills improved and they realized that they could do it all in less time. They reported a learning curve in their capacity to provide feedback and use the format efficiently. Teachers admitted that the format promoted more intensive interaction with residents.

When the resident doesn’t get it right, I have the opportunity to discuss with him different diagnostic or therapeutic alternatives for the patient... (Interview F. B. 4.2)

Feedback strategies
The teachers’ views with regard to feedback strategies related to the feasibility of constructive feedback.

(a) Feasibility of constructive feedback delivery
All teachers identified providing feedback as a key aspect of their role as educators, but they also indicated that the feedback they gave was generally brief and nonspecific. They also felt insecure because they had not been trained in feedback delivery and they worried that their feedback might cause frustration, loss of motivation, or diminution of self-esteem. Furthermore, they did not know how to respond when a resident showed a certain reaction, such as anxiety. The teachers expressed satisfaction about the opportunity to assess performance immediately after direct observation. They appreciated that the mini-CEX enabled them to give residents immediate, constructive, and structured feedback, including recommendations for improvement and action plans. It was the teachers’ experience that the mini-CEX facilitated the delivery of feedback and promoted more extensive and specific feedback. As noted before, they experienced a steep learning curve with frequent use of the format.

One can orient residents by explaining to them what their weak points are and how to improve them. Feedback can be provided in an organized way by following the different items to be assessed.... (Interviewee T 4.1)
Discussion

We conducted a qualitative study to explore clinical teachers’ experiences with the mini-CEX as an assessment tool, their perceptions of its feasibility and the influence of the format on their pass/fail decision strategies and feedback delivery. As for feasibility, the teachers thought the mini-CEX was easy to apply and integrate with daily patient care activities. The teachers’ views suggest that the mini-CEX has acceptable feasibility, and this resonates with the literature. Kogan et al. (2003) defined feasibility as the percentage of completed forms, average completion time and satisfaction ratings (Kogan, Bellini, & Shea, 2003). They analyzed data from 162 students that underwent mini-CEX evaluations and completed a total of 1,297 forms (89% completion rate), with a mean number of 7.9 forms per student (range: 2-10, median 8). Torre et al. (2007) considered one mini-CEX per month to be sufficient to determine its feasibility, and they achieved a 100% completion rate (Torre et al., 2007). As for strategies to reach pass/fail decisions, the teachers’ main response was that such decisions made them feel uncomfortable and uncertain, because they did not know what type of information or specific behaviors they should document during the mini-CEX. They also felt that contextual factors were likely to bias their judgment, and as a result they were likely to be too lenient. Other authors have reported similar findings. Dudek et al. (2005) explored factors, identified by supervisors as affecting their willingness to report poor clinical performance when completing In-Training Evaluation Reports (ITERs) (Dudek, Marks, & Regehr, 2005). They identified four major barriers to judgments of poor performance: 1. lack of documentation, 2. lack of knowledge as to what specifically to document, 3. anticipation of an appeal process and 4. lack of remediation options. In the present study teachers’ judgments appeared to be most strongly affected by the absence of well-defined standards and probably also by the high stakes for residents and the relationship between teacher and resident. The teachers in our study indicated that uncertainty surrounding pass/fail decisions could cause them to be overly lenient. Comparable problems with standards for pass/fail judgments have been reported by other authors. Littlefield et al. (1991) found that, compared with their colleagues, 27% of assessors were either very lenient or strict in assigning in-training evaluating scores (Littlefield et al., 1991). Apparently, there was considerable variation in the standards and expectations which assessors think should be met by third-year medical students. Turnbull et al. (2002) pointed to the dissonance between the roles of teacher and evaluator as a potential source of error when attending physicians evaluate the students they supervise (Turnbull & Van Barneveld, 2002). The work done by Magzoub et al. (1998) suggests that face-to-face evaluation could yield higher scores than any other type of evaluation not involving direct contact between assessor and assesseee at the time of the exam (Magzoub, Schmidt, Abdel-Hameed, Dolmans, & Mustafa, 1998). The findings with regard to pass/fail
decisions suggest that, when a new assessment format is introduced in a residency program, directors and faculty members should consider setting specific guidelines and performance standards for each level of training, thereby facilitating longitudinal assessment of individual residents. Alternatively, it might be wise to use individual mini-CEX evaluations as purely formative tools, which do not require a pass or fail decision, and to use the collection of all mini-CEX judgments for summative decisions. Such a collection of judgments, together with other assessment and performance information, might be a better basis for summative decision making. In this way, the emphasis in pass-fail decisions shifts to remediation and longitudinal personal development, which results in protection of both the individual assessor and the relationship with the resident. As a result, the focus in each separate mini-CEX evaluation is the resident’s clinical performance in a specific situation, and the teacher does not have to pronounce judgment as to whether the resident is a good doctor. Another issue that deserves special consideration is feedback delivery. In the present study the teachers saw the mini-CEX as an excellent opportunity to deliver instant feedback. One of the aims of assessment of residents’ performance based on observation of individual patient encounters is to promote and optimize feedback, which requires teachers to identify which aspects went well, which needed improvement, and what action should be taken. This type of feedback is likely to diminish the pressure on both assessor and learner. Moreover, the richness of the feedback in individual encounters can contribute to informed and defensible aggregate judgment across encounters. Several publications have described the mini-CEX as a tool for feedback delivery. Holmboe et al. (2004) reported results of an analysis of 107 feedback sessions after application of the mini-CEX: in 80% of the sessions the supervisor gave the residents advice regarding performance improvement at least once, in 61% the supervisor asked the resident to give his or her reactions, in 34% the supervisor asked the resident for self-observation and in 8% supervisor and resident developed an action plan together (Holmboe et al., 2004). As far as duration is concerned, Kogan et al. (2003) reported that the feedback session following the patient encounter of the mini-CEX lasted eight minutes and Hauer (2000) reported similar results (Hauer, 2000; Kogan et al., 2003). Later, Alves de Lima et al. (2007) analyzed 253 mini-CEX encounters and found an average duration of 17 minutes (Alves de Lima et al., 2007). Although it is important that an evaluation instrument should allow for narrative feedback, it is difficult to get teachers to provide written feedback. We therefore need strategies to facilitate written feedback through either technological support (i.e. voice recording or speech recognition, or electronic formats that aggregate scores and feedback across multiple assessments) or procedural measures (the learner writes down the verbal feedback). From the teachers’ views in the present study we can derive several practical recommendations to be considered in introducing the mini-CEX into a training program: (1) analyze with the teachers involved each of the competencies
to be assessed, (2) identify what is important to observe, (3) agree on minimum requirements for residents based on their levels of expertise/experience, (4) make sure that assessment forms are available in all the different locations where the mini-CEX encounter could take place (coronary care unit, emergency room) or, alternatively, design pocket-sized forms, (5) schedule the session with the resident (alternatively, residents can request the teacher to observe them), (6) observe residents' performance, (7) complete the form and (8) deliver feedback immediately after observation (Table 1).

### Table 1. Recommendations to enhance feasibility

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Analyze with teachers involved each of the competencies to assess</td>
</tr>
<tr>
<td>2</td>
<td>Identify what is important to observe</td>
</tr>
<tr>
<td>3</td>
<td>Agree on minimum requirements for each resident according to level of expertise/experience</td>
</tr>
<tr>
<td>4</td>
<td>Distribute the forms throughout the different observation scenarios or design pocket forms</td>
</tr>
<tr>
<td>5</td>
<td>Schedule the session with the resident. The resident may also request the teacher to be observed</td>
</tr>
<tr>
<td>6</td>
<td>Observe performance</td>
</tr>
<tr>
<td>7</td>
<td>Complete the form</td>
</tr>
<tr>
<td>8</td>
<td>Deliver immediate feedback</td>
</tr>
</tbody>
</table>

### Limitations of the study

The external validity of the study is limited because we investigated the teachers' reported experiences with the mini-CEX in a single residency program. Another limitation is that this particular mini-CEX required residents to perform to a high standard, with only one fail allowed per mini-CEX, which implied reassessment by the program director. The teachers' limited experience in using the format could be considered another limitation.

### Conclusions

Teachers perceive the mini-CEX as a useful formative assessment tool that promotes direct observation and facilitates instant feedback delivery. Implementation of teacher training programs and development of specific guidelines and performance standards for each level of residency training could be helpful in setting clear pass/fail thresholds. Also, interpersonal and institutional relationships and preconceptions regarding a resident's performance profile are relevant contextual factors, which can impact pass/fail decisions. An individual mini-CEX should be used primarily as a formative assessment tool, while the combined judgments on all mini-CEX assessments, together with other assessment and performance information, could be used for summative decision making. In this way, each individual
mini-CEX assesses clinical performance in this specific situation and the learner is not judged on his/her general qualities as a good doctor. Just as students and examinees adhere to understandable behavioral patterns, teachers display certain patterns of human behavior. Assessors often bring specific values to assessment, based on personal experiences, beliefs, and conceptions (or misconceptions). The central lesson is that, with instruments like the mini-CEX, assessors are not “passive measurement instruments”, but active judges who, implicitly or explicitly, bring interpretations and values to the assessment which can have considerable impact (Govaerts et al., 2007). The theoretical implication is that we need to further clarify the following: What do assessors bring to the assessment and why? How can they be supported? What contextual influences affect them in making judgments? Before we have more answers here, the practical implication is that we need to be cautious with high stakes summative mini-CEX implementations. In the meantime we should encourage its formative value and use.
References


Chapter 5

A laboratory study on the reliability estimations of the Mini-CEX

A. Alves de Lima, D. Conde, J. Costabel, J. Corso, CPM van der Vleuten

Advances Science Health Education 2013; 18 (1):5-13
Summary

Reliability estimations of workplace-based assessments with the mini-CEX are typically based on real-life data. Estimations are based on the assumption of local independence: the object of the measurement should not be influenced by the measurement itself and samples should be completely independent. This is difficult to achieve. Furthermore, the variance caused by the case/patient or by assessor is completely confounded. We have no idea how much each of these factors contribute to the noise in the measurement. The aim of this study was to use a controlled setup that overcomes these difficulties and to estimate the reproducibility of the mini-CEX. Three encounters were videotaped from 21 residents. The patients were the same for all residents. Each encounter was assessed by 3 assessors who assessed all encounters for all residents. This delivered a fully crossed (all random) two-facet generalizability design. A quarter of the total variance was associated with universe score variance (28%). The largest source of variance was the general error term (34%) followed by the main effect of assessors (18%). Generalizability coefficients indicated that an approximate sample of 9 encounters was needed assuming a single different assessor per encounter and assuming different cases per encounter (the usual situation in real practice), 4 encounters when 2 raters were used and 3 encounters when 3 raters are used. Unexplained general error and the leniency/stringency of assessors are the major causes for unreliability in mini-CEX. To optimize reliability rater training might have an effect.
Introduction

Reliability estimations of workplace-based assessments with the mini-CEX are typically based on real-life data obtained from different mini-CEX assessments, performed in clinical practice by different assessors on different occasions (Durning et al., 2002; Norcini et al., 2003; Alves de Lima et al., 2007). Ideally, reliability estimations are based on the assumption of local independence: the object of the measurement should not be influenced by the measurement itself and samples should be completely independent. With real life data this is difficult to achieve, however, since every mini-CEX will have a learning effect and, with one assessor and one case/patient per assessment, assessor variance and case variance are easily confounded, making it difficult to tease apart the effects of these different variables on the measurement. In a literature review of instruments for single-encounter workplace-based assessment, including the mini-CEX, Pelgrim found eight studies reporting reliability results, showing mostly acceptable (>0.8) reliability with a feasible sample size of ten encounters (Pelgrim et al., 2011).

These results were based on data collected in real-life settings. Apparently, in real life use, reliable assessment based on the mini-CEX requires input from numerous different assessors, a conclusion supported by Cook, who revealed a reproducibility coefficient of 0.23 (0.70 for 10 assessors or encounters) (D. A. Cook, Beckman, Mandrekar, & Pankratz, 2010). Hill, analyzed a total of 3400 mini-CEX forms and he found that reliability can be achieved by aggregating scores over 15 encounters and it was limited by variable examiner stringency (Hill, Kendall, Galbraith, & Crossley, 2009). Weller, collected 331 assessments from 61 trainees and 58 assessors. He also found that variable assessors stringency means that large numbers of assessors are required to produce reliable scores (Weller et al., 2009). Moreover, Kogan, who systematically reviewed the literature on tools for direct observation and assessment of clinical skills, found frequent reports of suboptimal inter-assessor reliability (<0.70) (Kogan, Holmboe, & Hauer, 2009). The only study to examine reliability in a controlled laboratory setting reported less favorable results from an analysis of ratings of a total of 48 cases by eight practicing physicians (Margolis et al., 2006). The practicing physicians were recruited from around the country. They were trained in a highly structured program to use the mini CEX rating form who each individually rated videotaped performances of ten examinees on six different cases from the Step 2 (clinical skills) examination of the United States Medical Licensing Exam, a standardized high stake test. The training program was divided in three different meetings. At each meeting, the training session lasted approximately three hours. Assessor variance turned out to be consistently larger than examinee variance, a finding suggesting that differences in assessor stringency contributed considerably more to the measurement error than did case specificity and supported by the difference between the low reliability coefficient (0.39) with
one assessor judging ten encounters and the considerably higher reliability coefficient (0.83) with ten assessors judging one encounter each (Margolis et al., 2006).

In order to overcome the drawbacks of real-life datasets, we designed a controlled setup with multiple assessors all individually assessing the same multiple cases. As a difference with Margolis study, we used standardized patients in the normal hospital setting where the residents demonstrated probably more habitual performance, where the raters were much less prepared and selected as in a high stakes setting. In other words, this study is probably more authentic to the usual mini-CEX in actual practice. A fully crossed design was used to investigate the variance components of the mini-CEX. For reliability indices of the mini-CEX, we used a fully nested design and a residents and assessors nested within cases design expecting that it would reveal comparable more detailed information on assessor-and case-related sources of variance in mini-CEX ratings in vivo conditions.

Methods

The study was conducted at the Cardiovascular Institute of Buenos Aires (ICBA), a 55-bed cardiovascular teaching hospital in the federal district of Buenos Aires province, Argentina. Both the institute and the cardiology residency program are affiliated with the University of Buenos Aires (UBA).

Participants

A total of 21 residents from each year of the four-year cardiology training program were invited to participate in the study: five residents from the first year, four residents from the second year, and six residents from both the third and the fourth year. Participation was voluntary and after explaining the purposes of the study to the residents, all of them agreed to participate. All the residents were videotaped during the same three encounters with three different simulated patients: a 53-year-old male presenting to the clinic seven days after an uncomplicated acute myocardial infarction, a 37-year-old dyslipidemic female attending the clinic for a blood pressure check-up and for blood results one week after an episode of high blood pressure (170/95 at the ER) for which the ER physician had recommended a low salt diet and regular exercise, and requested a lipid profile, and a 34-year-old male consulting for a preoperative cardiovascular risk evaluation prior to a scheduled laparoscopic cholecystectomy.

Each one of three internal medicine specialists from outside the institute who all had previous experience with the mini-CEX and were involved in medical education individually assessed all encounters of all participating residents. The following criterion was used to select candidates: faculty members who had used the
mini-CEX to assess residents on at least 10 occasions in their own internal medicine program.

Before the actual assessments, the specialists took part in a training session lasting approximately two hours. The assessors were invited to reflect on each of the domains of the mini-CEX and to discuss what was important to be observed, and what were the minimum performance requirements to be met.

Using a nine-point scale with 1, 2, and 3 indicating unsatisfactory performance, 4 marginal performance, 5 and 6 satisfactory performance, and 7, 8, and 9 superior performance, the assessors rated residents' performance on the competencies medical interviewing skills, physical examination skills, humanistic qualities, clinical judgment, counseling skills, organization skills and efficiency as well as on overall clinical competence. Total scores were calculated by averaging across the competencies (in line with Cook’s suggestion of uni-dimensionality) (Cook, Kallen, & Amtmaan, 2009).

The method we used offers a fully crossed (all random) two-facet (assessors and cases) generalizability design. The research protocol was ethically approved by the Institutional Review Board of the Instituto Cardiovascular de Buenos Aires

Analysis

For each resident we averaged scores across items of the Mini-CEx rating scale, leading to a case score. Descriptive statistics were calculated per each case, for each assessor, for the overall case across assessors and the total scores for all cases. Variance component estimations were performed for each of the seven sources of variance associated with a fully crossed design. For the D-studies (estimating the reliability indices) we used two different designs: a fully nested design and a design with residents nested within cases and crossed with assessors. In the fully nested design, residents and assessors were nested within cases, because this would enable comparison of our dataset with in vivo datasets representing different cases (patients) and different assessors (Crossley et al., 2002). In some in vivo conditions, however, there may be only one assessor available for all the residents in the setting, and consequently cases are nested within residents but not within assessors. All analyses were conducted using the mGENOVA software package.

Results

Table 1 shows the mean scores and standard deviations for each case, for each assessor, for the overall case score across assessors, and for the total score for all cases.
The estimated variance components for all potential sources of variance (Table 2) showed that the general error term (Vrca) (0.58, 34%) and systematic assessor stringency/leniency (a) (0.31, 18%) are the main sources of variance. The other assessor-related variance components accounted for relatively small percentages of the variance with Vcr 0.16 (9%) or case specificity, Var 0.12 (7%) and Vca 0.07 (4%). Table 3 shows the reliability coefficients for the fully nested design. With one single assessor for one encounter - but different ones for different encounters - the usual situation in residency training, approximately nine encounters would be needed to achieve a reliability of 0.8, with substantially fewer encounters required as more assessors are added, the required number dropping to as low as four encounters with two assessors and even further to only three encounters with three assessors.

Table 4 presents the reliability coefficients when the same assessor is used across all encounters. To achieve a reliability of 0.80 more than fifteen encounters would be needed.
Table 1. Means scores and standard deviations for the 3 cases split up per assessors, the overall scores for all assessors per case and the total score for all cases

<table>
<thead>
<tr>
<th></th>
<th>Medical Interviewing skills</th>
<th>Physical examination skills</th>
<th>Humanistic qualities/professionalism</th>
<th>Clinical Judgement</th>
<th>Counselling skills</th>
<th>Organization/efficiency</th>
<th>Global clinical competence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1 A1</td>
<td>M 6.61</td>
<td>6.9</td>
<td>7.61</td>
<td>6.09</td>
<td>6.14</td>
<td>6.14</td>
<td>6.04</td>
</tr>
<tr>
<td></td>
<td>SD 1.85</td>
<td>1.25</td>
<td>0.58</td>
<td>2.09</td>
<td>2.05</td>
<td>1.79</td>
<td>1.68</td>
</tr>
<tr>
<td></td>
<td>A2 M 6.14</td>
<td>6</td>
<td>6.19</td>
<td>6.3</td>
<td>6.38</td>
<td>6.23</td>
<td>6.57</td>
</tr>
<tr>
<td></td>
<td>SD 1.45</td>
<td>0.5</td>
<td>1.5</td>
<td>1.34</td>
<td>1.85</td>
<td>1.6</td>
<td>1.59</td>
</tr>
<tr>
<td></td>
<td>A3 M 5.95</td>
<td>6.11</td>
<td>6.04</td>
<td>2.96</td>
<td>5.71</td>
<td>6</td>
<td>5.42</td>
</tr>
<tr>
<td></td>
<td>SD 1.85</td>
<td>1.28</td>
<td>1.8</td>
<td>1.72</td>
<td>2.14</td>
<td>1.78</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>O M 6.24</td>
<td>6.25</td>
<td>6.62</td>
<td>6.06</td>
<td>6.08</td>
<td>6.13</td>
<td>6.02</td>
</tr>
<tr>
<td></td>
<td>SD 1.79</td>
<td>1.22</td>
<td>1.54</td>
<td>1.73</td>
<td>2.01</td>
<td>1.7</td>
<td>1.73</td>
</tr>
<tr>
<td>Case 2 A1</td>
<td>M 7.33</td>
<td>7</td>
<td>7.66</td>
<td>6.52</td>
<td>6.9</td>
<td>6.57</td>
<td>6.52</td>
</tr>
<tr>
<td></td>
<td>SD 1.15</td>
<td>1.22</td>
<td>0.57</td>
<td>2.33</td>
<td>1.17</td>
<td>1.2</td>
<td>1.32</td>
</tr>
<tr>
<td></td>
<td>A2 M 6.09</td>
<td>5.95</td>
<td>5.71</td>
<td>5.8</td>
<td>5.9</td>
<td>5.95</td>
<td>6.19</td>
</tr>
<tr>
<td></td>
<td>SD 1.13</td>
<td>0.38</td>
<td>1</td>
<td>0.81</td>
<td>1.41</td>
<td>0.97</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>A3 M 4.9</td>
<td>6.23</td>
<td>5.76</td>
<td>5.04</td>
<td>5</td>
<td>5.66</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>SD 1.48</td>
<td>1.3</td>
<td>1.3</td>
<td>1.43</td>
<td>1.3</td>
<td>1.31</td>
<td>1.24</td>
</tr>
<tr>
<td></td>
<td>O M 6.11</td>
<td>6.4</td>
<td>6.38</td>
<td>5.79</td>
<td>5.81</td>
<td>6.06</td>
<td>5.84</td>
</tr>
<tr>
<td></td>
<td>SD 1.59</td>
<td>1.12</td>
<td>1.34</td>
<td>1.73</td>
<td>1.65</td>
<td>1.21</td>
<td>1.42</td>
</tr>
<tr>
<td></td>
<td>SD 1.96</td>
<td>0.87</td>
<td>0.49</td>
<td>1.85</td>
<td>2.34</td>
<td>1.69</td>
<td>1.62</td>
</tr>
<tr>
<td></td>
<td>A2 M 5.71</td>
<td>6</td>
<td>5.71</td>
<td>5.9</td>
<td>5.33</td>
<td>5.76</td>
<td>5.71</td>
</tr>
<tr>
<td></td>
<td>SD 0.84</td>
<td>0</td>
<td>0.9</td>
<td>0.53</td>
<td>1.35</td>
<td>0.53</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>A3 M 5.52</td>
<td>5.61</td>
<td>6.14</td>
<td>5.38</td>
<td>4.71</td>
<td>5.23</td>
<td>5.04</td>
</tr>
<tr>
<td></td>
<td>SD 1.36</td>
<td>0.97</td>
<td>1.15</td>
<td>1.49</td>
<td>1.84</td>
<td>1.67</td>
<td>1.65</td>
</tr>
<tr>
<td></td>
<td>O M 6.06</td>
<td>6.37</td>
<td>6.49</td>
<td>5.81</td>
<td>5.38</td>
<td>5.81</td>
<td>5.79</td>
</tr>
<tr>
<td></td>
<td>SD 1.57</td>
<td>1.09</td>
<td>1.2</td>
<td>1.63</td>
<td>1.94</td>
<td>1.46</td>
<td>1.53</td>
</tr>
<tr>
<td>All cases Total</td>
<td>M 6.14</td>
<td>6.34</td>
<td>6.5</td>
<td>5.92</td>
<td>5.8</td>
<td>6</td>
<td>5.88</td>
</tr>
<tr>
<td>score</td>
<td>SD 1.62</td>
<td>1.14</td>
<td>1.37</td>
<td>1.64</td>
<td>1.84</td>
<td>1.47</td>
<td>1.56</td>
</tr>
</tbody>
</table>

A1 = assessor 1, A2 = assessor 2, A3 = assessor 3, O = overall score, M = mean, SD = Standard deviation
Table 2. Estimated variance components, standard errors, and relative size of variance components

<table>
<thead>
<tr>
<th>Source of variance</th>
<th>Explanation</th>
<th>Estimated variance components</th>
<th>Standard Error</th>
<th>% of total variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vr</td>
<td>Systematic variability of residents</td>
<td>0.48431</td>
<td>0.19709</td>
<td>28</td>
</tr>
<tr>
<td>Vc</td>
<td>Systematic variability of cases (case difficulty)</td>
<td>0.00000</td>
<td>0.02150</td>
<td>0</td>
</tr>
<tr>
<td>Va</td>
<td>Systematic variability of assessors (leniency/stringency)</td>
<td>0.30925</td>
<td>0.24682</td>
<td>18</td>
</tr>
<tr>
<td>Vcr</td>
<td>Variability of residents across cases</td>
<td>0.15974</td>
<td>0.08302</td>
<td>9</td>
</tr>
<tr>
<td>Var</td>
<td>Assessor variability for some residents</td>
<td>0.12108</td>
<td>0.07523</td>
<td>7</td>
</tr>
<tr>
<td>Vca</td>
<td>Assessor variability for some cases</td>
<td>0.07113</td>
<td>0.05726</td>
<td>4</td>
</tr>
<tr>
<td>Vrca</td>
<td>General error term</td>
<td>0.58305</td>
<td>0.09106</td>
<td>34</td>
</tr>
<tr>
<td>Σ</td>
<td></td>
<td>1.72855</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Reliability estimates as a function of the number of cases and assessors for the situation where residents are given different cases with different assessors

<table>
<thead>
<tr>
<th>Number of cases</th>
<th>One assessor per case</th>
<th>Two assessors per case</th>
<th>Three assessors per case</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.33</td>
<td>0.49</td>
<td>0.59</td>
</tr>
<tr>
<td>2</td>
<td>0.49</td>
<td>0.66</td>
<td>0.74</td>
</tr>
<tr>
<td>3</td>
<td>0.59</td>
<td>0.74</td>
<td>0.81</td>
</tr>
<tr>
<td>4</td>
<td>0.66</td>
<td>0.80</td>
<td>0.85</td>
</tr>
<tr>
<td>5</td>
<td>0.71</td>
<td>0.83</td>
<td>0.88</td>
</tr>
<tr>
<td>7</td>
<td>0.77</td>
<td>0.87</td>
<td>0.91</td>
</tr>
<tr>
<td>9</td>
<td>0.81</td>
<td>0.90</td>
<td>0.93</td>
</tr>
<tr>
<td>11</td>
<td>0.84</td>
<td>0.91</td>
<td>0.94</td>
</tr>
<tr>
<td>13</td>
<td>0.86</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td>15</td>
<td>0.88</td>
<td>0.94</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Table 4. Reliability estimates as a function of the number of cases and assessors for the situation where residents are given different cases but with the same assessors.

<table>
<thead>
<tr>
<th>Number of cases</th>
<th>One assessor for all cases</th>
<th>The same two assessors for all cases</th>
<th>The same three assessors for all cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.36</td>
<td>0.49</td>
<td>0.56</td>
</tr>
<tr>
<td>2</td>
<td>0.50</td>
<td>0.64</td>
<td>0.70</td>
</tr>
<tr>
<td>3</td>
<td>0.58</td>
<td>0.71</td>
<td>0.77</td>
</tr>
<tr>
<td>4</td>
<td>0.63</td>
<td>0.75</td>
<td>0.80</td>
</tr>
<tr>
<td>5</td>
<td>0.66</td>
<td>0.78</td>
<td>0.83</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>0.81</td>
<td>0.86</td>
</tr>
<tr>
<td>9</td>
<td>0.73</td>
<td>0.83</td>
<td>0.87</td>
</tr>
<tr>
<td>11</td>
<td>0.75</td>
<td>0.85</td>
<td>0.88</td>
</tr>
<tr>
<td>13</td>
<td>0.76</td>
<td>0.85</td>
<td>0.89</td>
</tr>
<tr>
<td>15</td>
<td>0.77</td>
<td>0.86</td>
<td>0.90</td>
</tr>
</tbody>
</table>

66
Discussion

We examined assessment data in a fully crossed design in which every resident was assessed by the same three assessors to assess performance on the same cases using the mini-CEX, a design that allows for the most efficient variance component analysis, but nevertheless is fairly uncommon. The results give rise to two main conclusions: unexplained general error and assessor leniency/stringency (systematic across assessors) appear to be the major causes of unreliability of mini-CEX assessments.

Within the univariate framework, several results are of interest. The small examinee by case variance (9%) appears to indicate a small effect of content specificity, while the relatively large examinee variance, which was consistently larger than assessor variance, suggests that inter-rater differences in stringency make a considerably larger contribution to measurement error than does case specificity. This seems quite surprising, since in standardized testing, like OSCE, the reverse is generally reported (high content specificity, lower assessor specificity). It may be the case that in realistic settings expert judges assess something that is quite generalizable across cases, but at the same time - probably due to the unstandardized and global nature of the judgment - inherently susceptible to rater effects.

The generalizability coefficients that we found indicate that a sample of approximately nine encounters would suffice with one assessor per encounter but different assessors for different encounters, while fifteen encounters would be needed when there is only one single assessor for all encounters. Having more than one assessor per encounter - an extremely rare situation in real practice - resulted in a substantial reduction of the number of encounters needed, with two assessors halving the number of encounters required. Apparently, case and assessor variance have similar effects on measurements obtained with the mini-CEX.

The results appear to be consistent with the literature. Margolis et al. (2006) also found that differences in assessor stringency made the greatest contribution to measurement error, and that a higher number of assessors for each examinee could enhance score stability, even with fewer cases (Margolis et al., 2006). Similar assessor effects were found by Weller, based on analysis of data from 331 assessment forms, 61 trainees, and 58 assessors, revealing variance of assessor stringency to be the main cause of unreliability, contributing 40% to score variation (Weller et al., 2009). In an analysis of a total of 3,499 mini-CEX forms, Hill found a considerable contribution (29%) of assessor stringency to the score variation as well, from which they inferred some practical implications (Hill et al., 2009). They suggested that there might be some value in assessor training or selection, since stringency variation tells us something about the internalized standards of the assessors. Consequently, it seems advisable to promote uniformity of standards through assessor training by defining what is important for assessors to observe as well as the mini-
mum requirements for resident performance at different levels of expertise/experience, and also by discussing rating decisions. However, since sampling across several assessors may be equally effective in ameliorating the effect of stringency variation, the authors also proposed what they called a crossed assessment design, in which each trainee is assessed by the same group of assessors ensuring that stringency variations are evenly distributed across trainees, and consequently no-one is unduly disadvantaged (Hill et al., 2009).

There are limitations to this study due to specific characteristics of the dataset. First of all, the small sample size and the resulting precision of variance component estimation (as can be seen from the standard errors in table 2) diminish the generalizability of the findings to the typical operational application of the mini-CEX. There are additional limitations relating to differences between the conditions of the study and those of the mini-CEX in clinical practice: the use of videotaped performance rather than direct observation, information about diagnosis and management plan being obtained in a written format as opposed to face-to-face interview, the use of standardized patients, the two-hour training session for the assessors exceeding the usual exposure of assessors to such training, and the assessors not knowing the residents whose performance they judged, while in real practice the assessor resident relationship tends to inflate scores. Other limitations are that residents were from different years and consequently differed in expertise, which may have inflated the variance components of the residents.

This study addresses reliability issues derived from standardized but highly realistic assessment setting. We used standardized patients in a normal hospital setting where residents show their habitual performance and where the raters were less trained as in a previous similar laboratory controlled study (Margolis et al., 2006). In other words, our laboratory setting has more ecological validity. This study adds information about reliability issues closer to the real world. There are two main implications for practice. First, regarding the performance based assessment, the value of the assessment appears to be determined by the users of the instruments rather than by the instruments themselves (van der Vleuten et al., 2010). We agree with Hill on the fact that assessors training might in some way be helpful towards optimizing reliability (Hill et al., 2009). The understanding of the factors impacting on assessor’s judgments’ and ratings after direct observation is crucial and should be taken into account at the time of organizing the assessor training sessions. Kogan identified four primary themes that provide insights into the variability of assessors’ assessment of residents’ performance: the frame of reference used by assessors when translating observation into judgments and rating, the level of inferences that are used during the direct observation process, the methods by which judgments are synthesized into numerical ratings and the contextual factors (Kogan, Conforti, Bernabeo, Iobst, & Holmboe, 2011). Second, in clinical practice where only one assessor is available, multiple observations are the key
for reliable scores. The required sample size of approximately nine mini-CEX assessments that emerged from this study is in accordance with estimations based on actual-life data.
References


Chapter 6

Educational impact of the mini-CEX on medical students’ performance

A. Alves de Lima & CPM van der Vleuten

This paper was accepted in a resumed version in Medicina (Bs As), 2013, 73(3): 286-291
Summary

Introduction: Validity, reliability and acceptability of the Mini-CEX utility have been studied over the past few years, but there is still relatively little knowledge about its educational impact or outcome. The purpose of this study was to measure changes in clinical skills proficiency in a cohort of fourth-year medical students assessed during their clerkship with the mini-CEX.

Methods: A random cluster sampling study was designed to assess behavioral changes. The students were divided in two consecutive groups. All the groups received the same contents and teaching strategies. The difference was that group 1 was assessed with 3 formative Mini-CEX during their clerkship period and group 2 was not. At the end of the three weeks each group was summatively assessed with the Mini-CEX with a simulated patient observed by two assessors.

Results: 50 fourth-year medical students were enrolled in the study. Group 1 was constituted by 27 students and Group 2 by 23 students. There were no significant differences in the means scores between the groups in any of the domains. (Interviewing skills, physical examination, professionalism, clinical judgment, organization and global clinical competence)

Discussion: We could not find significant differences in clinical performance and behavior between the groups. The low number of encounters students had, the short time they were exposed to the mini-CEX may have had an influence on these results.

Conclusion: We might need to move to an interventionist, experimental model to establish whether the workplace-based assessment makes a difference on learning outcome.
Introduction

The Mini-CEX is a workplace-based assessment tool that intends to evaluate students at the "Does" level of the Miller’s pyramid, that is, in actual situations and settings (Norcini et al., 1995). It has been designed to incorporate both the skills that students require in actual patient encounters, along with the educational interactions that supervisors routinely encounter with students during the teaching rounds in their clerkships. It is an evaluation method that promotes the assessment of clinical skills along with attitudes and behaviors that are essential in high-quality patient care. It is a performance-based evaluation method that is used to assess selected clinical competencies (e.g. patient charts and physical examination, also communication and interpersonal skills) in the medical training context. A single faculty member observes and evaluates a student while he conducts a thorough history and physical examination on an 'in' or outpatient or a patient in the emergency department. After asking the students for a diagnosis and treatment plan, the supervisor completes a short evaluation form and gives direct feedback. As the interaction is relatively brief and occurs as a natural part of the process in the training environment, each individual can be evaluated on several occasions and by various supervisors. Certain aspects of Mini-CEX’s utility, particularly its feasibility, reliability and validity, have been studied over the past few years but information related to behavioral changes is weak and scarce. Pelgrim reviewed the literature on instruments for work-based assessments in single clinical encounters such as the mini-CEX and they have reached the following conclusions: first, they found that feasibility is generally deemed good and that although assessor training was found to be crucial for a successful implementation of the assessment it occurs sparsely and second, validity is supported by significant and strong correlations with other valid assessment instruments (Pelgrim et al., 2011). However, they reported that the evidence obtained from the few studies that have been carried out on educational impact is not yet conclusive.

Kogan have evaluated the educational impact of the Mini-CEX in medical students (Kogan et al., 2003). Using a nine-point scale (1=low and 9=high), they find that satisfaction with the exercise was rated as 6.8 by students. Torre have also investigated the perceptions of the Mini-CEX in medical students, but this time with a PDA-based mini-CEX (Torre et al., 2007). PDA-based mini-CEX was highly rated by students as a valuable technology based tool to document direct supervision of clinical skills. Hill studied the feasibility, and the acceptability of the Mini-CEX in a group of undergraduate medical students at Southampton University. They found that medical students were satisfied with the exercise because it reflects their true abilities (Hill et al., 2009). Undoubtedly, the assessment system constitutes the most vital factor influencing students learning behavior (Gilliers, Schuwirth, Hermann, Adendorff, & van der Vleuten, 2012; Newble et al., 1990; Newble & Jaeger,
We wanted to study whether a formative assessment measure from the highest level of the pyramid like the mini-CEX can produce changes in learners’ behavior or clinical skills. Any assessment process should have educational impact. Considering the emphasis placed on the Mini-CEX as a method of formative performance assessment, so far the literature on learning impact has been primarily focused on perceptions and not on outcome measures of performance improvement (Miller & Archer, 2010).

The purpose of this study was to measure changes in clinical skills in a cohort of fourth-year medical students assessed during their clerkship with formative mini-CEX encounters. Our hypothesis is that the students who are assessed with formative Mini-CEX encounters during their cardiology clerkship will have a better clinical performance when faced with an unexpected simulated clinical situation at the end of the rotation.

**Methods**

The study was conducted at a 55-bed cardiovascular teaching hospital. Research subjects were a cohort of 50 fourth-year medical students participating in a 3-week cardiology clerkship during 2011. The learning objectives of the clerkship were to become familiar with all major aspects of the cardiovascular disease such as prevention, physical examination of the cardiovascular system, evaluation and treatment strategies, pharmacologic agents critical to treatment of cardiovascular disease, performance and interpretation of graded exercise stress testing (GXT) epidemiology and pathophysiology, in order to acquire improved skills in electrocardiogram interpretation and to gain exposure to cardiovascular imaging techniques and procedures. It was expected that all the students regularly attend the daily Internal Medicine morning report plus academic conferences as deemed appropriate by the rotation preceptor.

A random cluster sampling study was designed to measure behavioral changes. The students were divided randomly in two consecutive groups: All the groups had the same learning objectives and were trained on the same teaching strategies. The difference between the groups was that in Group 1 (G1) the participants were assessed with 3 Formative Mini-CEX during their clerkship period and Group 2 (G2) was not exposed to the formative Mini-CEX during the clerkship period. Practical circumstances made it impossible to perform a greater number of mini-CEX encounters. At the end of the three weeks each group was summatively assessed with the Mini-CEX with a single simulated patient. The competencies evaluated with mini-CEX were defined as follows: Medical Interviewing Skills: Facilitates the patient’s story-telling through effective use of questions/directions in order to obtain accurate and required information; responds appropriately to affect and non-verbal
cues, Physical Examination Skills: Shows efficiency and a logical sequence; balances screening/diagnostic steps towards problem; informs the patient and is sensitive to the patient’s comfort and modesty, Humanistic Qualities/ Professionalism: Shows respect, compassion, empathy, establishes trust; attends to the patient’s needs for comfort, modesty, confidentiality and information, Clinical Judgment: Selectively orders/perform appropriate diagnostic investigations/tests, considers risks and benefits, Organization/ Efficiency Skills: Prioritizes actions; uses time efficiently; is succinct and Overall Clinical Competence: Demonstrates judgment, synthesis, caring, effectiveness, and efficiency.

Performance was rated on a nine-point scale where 1, 2 and 3 indicated unsatisfactory performance, 4 marginal performance, 5 and 6 satisfactory performance, and 7, 8 and 9 superior performance. There were three different simulated patients: the first one, a 53-year-old male turning up at the clinic seven days after an uncomplicated acute myocardial infarction, the second, a 37-year-old dyslipidemic female attending the clinic for a blood pressure check-up and for blood results one week after an episode of high blood pressure (170/95 at the ER) for which the ER physician had recommended a low salt diet and regular exercise, and requested a lipid profile, and the third one, a 34-year-old male consulting for a preoperative cardiovascular risk evaluation prior to a scheduled laparoscopic cholecystectomy.

All the students knew they were going to be assessed at the end of the clerkship with a simulated patient (SP) but they could not know the problem of the SP they would be exposed to. All the assessors were cardiologists from the institute who had previous experience with the mini-CEX. Before the intervention, all assessors took part in a training session lasting approximately one hour. In the training session the assessors discussed and agreed on the issues to be observed and the mini-CEX assessment form to be met, until general consensus was reached.

For formative Mini-CEX encounters (G1) a single assessor observed and evaluated the students during an encounter with an “in” or “out”-patient or with a patient in the emergency department, while the residents conducted a history and performed a physical examination. After asking the students for diagnosis and treatment plan, the faculty member completed the Mini-CEX evaluation form and gave direct feedback. Mean Mini-CEX time was 22.77 minutes (SD 9.97).

For the summative mini-CEX encounters two assessors observed and evaluated the students during an encounter with one simulated patient. They were blinded as to whether the students had been assessed or not with the formative mini-CEX during the clerkship period. After asking the students for diagnosis and treatment plan, the faculty member completed the Mini-CEX form and gave direct feedback. Mean Mini-CEX time was 26.50 minutes (SD 8.55). If the student failed (rating 1, 2 or 3 on any domain) they had to schedule with the assessors a new encounter to repeat the Mini-CEX. This research protocol received ethical approval by the institutional Review Board of the Instituto Cardiovascular de Buenos Aires.
Score means for all the domains as well as on overall clinical competence from G1 (n=27) were compared with the means of G2 (n=23) using Mann-Whitney/Wilcoxon Two-Sample Test (Kruskal-Wallis test for two groups). Statistical significance was tested at the level of p value <0.05 (Table 1). All analyses were conducted using the SPSS version 11.0

Results

From February 2011 to May 2011, 50 fourth-year medical students were enrolled in the study. Sixty-five percent were women and their average age was 22.5. G1 was constituted by 27 students and G2 by 23 students. There were no baseline characteristic differences between both groups. Table 1 shows the means scores of each domain between G1 and G2. As seen in table 1, there were no significant differences between the groups in any of the domains

<table>
<thead>
<tr>
<th>Domains</th>
<th>Group 1 (n=27) Mean scores</th>
<th>Group 2 (n=23) Mean scores</th>
<th>p value (Kruskal-Wallis test for two groups)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interviewing skills</td>
<td>7.04 (SD= 1.22)</td>
<td>6.87 (SD= 1.17)</td>
<td>0.68</td>
</tr>
<tr>
<td>Physical examination</td>
<td>6.55 (SD= 1.25)</td>
<td>6.76 (SD= 1.04)</td>
<td>0.20</td>
</tr>
<tr>
<td>Professionalism</td>
<td>7.00 (SD= 1.10)</td>
<td>7.08 (SD= 0.90)</td>
<td>0.82</td>
</tr>
<tr>
<td>Clinical judgment</td>
<td>6.96 (SD= 1.28)</td>
<td>6.72 (SD= 1.07)</td>
<td>0.30</td>
</tr>
<tr>
<td>Organization</td>
<td>6.85 (SD= 1.35)</td>
<td>6.87 (SD= 1.14)</td>
<td>0.81</td>
</tr>
<tr>
<td>Global competence</td>
<td>6.81 (SD= 1.00)</td>
<td>6.87 (SD= 0.96)</td>
<td>0.94</td>
</tr>
</tbody>
</table>

SD: standard deviation, Group 1: students exposed to the formative Mini-CEX during the clerkship, Group 2: students NOT exposed to the formative mini-CEX during the clerkship.

Discussion

The purpose of this study was to measure in a more direct way, changes in clinical skills in a cohort of fourth-year medical students assessed during their clerkship with a formative mini-CEX intervention. Contrary to our expectations we could not find significant differences in clinical performance and behavior between the groups.

These results are aligned to the published data. The studies examining the mini-CEX showed largely positive results in terms of learners’ satisfaction but could not show changes in attitudes, skills or behaviors. The studies revealed that some house officers felt it could improve their clinical skills, but this evidence has not
been captured objectively and particularly numbers were small. A previous systematic review published by Kogan et al., investigating tools for direct observation and assessment of clinical skills found that few studies measure trainees’ attitudes during the mini-CEX encounters and none of them demonstrated improvements on clinical skills or patient care quality (Kogan et al., 2009). A similar conclusion was reached by Pelgrim (Pelgrim et al., 2011). They found that outcomes such as learning behavior, transfer of skills to new situations or improvement of patient care are not well investigated, although they are crucial for the evaluation of the educational impact.

There may be several explanations to these results. First, the intervention may not have been powerful enough since students were assessed in a low number of formative Mini-CEX encounters. Second, the performance of only 1 encounter with different SPs observed by two independent assessors for different students for the outcome measurement may have been insufficient. Third, another alternative may be a ceiling effect. All ratings were quiet high so the independent variable may have no longer an effect on a dependent variable.

Evidence on educational is lacking. No studies examined whether instruments improve learning, clinical skills or the quality of patient care. Given that the nature of the Mini-CEX is clearly formative, effects on learning and performance are the prime objective of this type of assessment. Existing research typically evaluates perception users, and although the outcomes are overwhelmingly positive, they do not provide compelling evidence of learning effects. More rigorous research will have to elucidate the educational effects of clinical work-based assessment. After utility of a tool has been demonstrated and guidelines for implementation developed, randomized studies designs should follow whenever possible to assess whether the tool affects educational outcomes. More multi institutional studies could help improve generalisability of the findings.
References


Chapter 7
Understanding raters errors: Reliability estimations of the mini-CEX using traditional and construct-aligned scales

A. Alves de Lima, A. Lavalle Cobo, F. Castro, L. Forti, M. Albertal, CPM van der Vleuten

Submitted
Abstract

Recently, Crossley have demonstrated that in real life settings Mini-CEX scales constructed to reflect the development of clinical sophistication and independence (CS) have higher utility than the traditional ones (IS), since they are more reliable and therefore raise the evidence of greater validity. The aim of this study is to reproduce these findings in a controlled setup and to evaluate the different variance components in both scales. Three encounters were videotaped from 21 residents (R). The patients were the same for all R. Each encounter was assessed by 3 assessors (A) who assessed all encounters for all R. The A assessed the encounters twice. The first time they assessed the encounters using the IS and 30 days later with the CS. Each A was an internal medicine specialist from outside the institute and was blinded to the level of expertise of the R. All of them had previous experience with the mini-CEX and were involved in medical education. This delivered a fully crossed (all random) two-facet generalisability design each time. For both scales, a third of the total variance was associated with universe score variance, IS: 36% vs CS 29%. The largest source of variance in the IS was of general error (49%), followed by the main effect of assessors (7%). In the CS the largest source of variance was of general error (34%) followed by the assessors’ variability for some residents (23%). Generalisability coefficients indicated that for both types of scales an approximate sample of 7 encounters was needed, assuming both the presence of one different assessor per encounter and the presentation of different cases per encounter (the usual situation in real practice): 4 encounters when 2 raters were used and 3 encounters in case 3 raters were used. According to the results obtained and contrary to our expectations the IS and the CS showed similar performance in terms of sources of variance and in the resulting reliability. Unexplained general error appears to be the major cause of unreliability of both scales followed by the assessor leniency/stringency in the IS and the assessors’ variability for some residents in the CS. The explanation for these results may be that assessors were blinded to the level of expertise of the residents. Probably, the knowledge of the level of training of the residents by the assessors could be important in the CS to consolidate a unique frame of reference during observation and rating between assessors to enhance its reliability. At least in our study, the Crossley’s original findings did not replicate, so it will be important to explore the differences in the contexts and potential limitations on the generalizability of his results. The phenomenon of construct alignment appears to be less straightforward than might be inferred from Crossley’s initial work, so further exploration of when and where and how it plays a role in improving assessments will be important for future research to establish.
Introduction

Understanding of the factors impacting on assessors' judgements and ratings after direct observation is crucial and should be taken into account at the time of implementing the mini-CEX (Alves de Lima et al., 2013). In fact, Govaerts et al. (2011) reported that raters make and justify judgements based on personal theories and personal constructs and their information processing seems to be affected by expertise differences (Govaerts, Schuwirth, van der Vleuten, & Muijtjens, 2011). Additionally, Gingerich et al. (2011) critically reviewed the literature on raters' idiosyncrasy on impression formation and they observed that raters may form categorical judgements about ratees as part of impression formation (Gingerich, Regehr, & Eva, 2011). In this direction, an assessment system requiring ordinal or interval ratings may inadvertently introduce conversion errors due to translation techniques unique to each rater. Accordingly, Crossley et al. (2011) have demonstrated that, in real life settings, mini-CEX scales constructed with categorical judgments that reflect the constructs of development of clinical sophistication and independence or entrustability have higher reliability than ordinal or interval ratings scales (Crossley, Johnson, Booth, & Wade, 2011). The construct of clinical sophistication, was well described by Pangaro (1999). He developed the RIME scheme, a synthetic framework that emphasis progressively expectations as a student progress as reporter, interpreter, manager/educator towards eventual independence through the clinical years and through the residency (Pangaro, 1999). In relation with the construct of independence or entrustability, ten Cate (2006) argues that clinical supervisors' judgments focus on the concept of entrustability. Entrusting a critical activity should led to the trainee being granted responsibility in all similar future circumstances. Once sound feedback has confirmed critical number of times that all went well, the entrustment could be formalized and considered a qualification to act independently (ten Cate, 2006). For this study, the original mini-CEX form that uses ordinal or interval scales (IS) was retained, but a second scale was added. In the new scale predetermined training levels anchors were accompanied by behavioral descriptors aligned to the construct of developing clinical sophistication and independence (CS). For example the original form includes IS ranging from well below expectations for stage of training to well above expectations for stage of training and in the new scale anchors included "demonstrates sound consultation skills, resulting in a sound history and/or examination findings. Shows basic clinical judgment following encounter". By comparing the IS scales and the CS scales used in parallel to assess 1843 medical trainee by the mini-CEX, they evaluated how scores reflect assessors' disagreement with one another and estimated reliability using generalisability theory. They found that CS substantially reduced assessor disagreement. The aim of this study is to reproduce these findings in a laboratory controlled set-up and to evaluate different variance components in both scales.
Methods

The study was conducted at the Instituto Cardiovascular de Buenos Aires (ICBA), a 70-bed cardiovascular teaching hospital with a residency program affiliated with the University of Buenos Aires (UBA) in the Autonomous City of Buenos Aires, Argentina.

Participants

Three individual meetings were held at the offices of the educational department during March 2012. At each meeting one assessor was trained during 1 hour to rate performances from a set on 63 videotaped clinical encounters with the mini-CEX. During each meeting each assessors was invited to reflect on each of the domains of the IS mini-CEX scale and on the anchors of the CS mini-CEX scales. They were also request to discuss what was important to observe and what were the minimum requirements to meet with each of the mini-CEX scales. The set of the 63 encounters were recorded previously for training and institutional research purposes but not specifically for this study. In that set 21 residents from each year of a 4-year training program; five post graduate year (PGY) 1 residents, four PGY-2 residents, six PGY-3 residents and six PGY-4 residents were videotaped during the same three encounters with three different simulates patients: a 53-year-old man presenting to the clinic seven days after an uncomplicated acute myocardial infarction. The second was a 37-year-old dyslipidemic woman who had been seen at the ER with an episode of hypertension (170/95) one week before; the ER physician had recommended a low salt diet and regular exercise, and requested a lipid profile; she came to the clinic for blood pressure check-up and interpretation of lab tests results. The last patient was a 34-year-old man consulting for a preoperative cardiovascular risk evaluation before a scheduled laparoscopic cholecystectomy. The three assessors trained were internal medicine specialists from outside the institute, had previous experience with the mini-CEX and were involved in medical education assessed individually all the video-taped encounters twice. The IS was used for the first evaluation and 30 days later the CS was used. All the assessors were blinded to the level of expertise of the residents. The following criterion was used to select the assessors: faculty members who had used the mini-CEX to assess residents on at least 10 occasions in their own internal medicine program. The IS was constructed using a nine-point scale with 1, 2, and 3 indicating unsatisfactory performance, 4 marginal performance, 5 and 6 satisfactory performance, and 7, 8, and 9 superior performance (Appendix 1).
The assessors rated residents’ performance on the competencies medical interviewing skills, physical examination skills, humanistic qualities, clinical judgment, counseling skills, organization skills and efficiency as well as on overall clinical competence. Total scores were calculated by averaging across the competencies (in line with Cook's suggestion of uni-dimensionality) (Cook et al. 2009). The CS was
constructed using predetermined training levels anchors that were accompanied by
behavioral descriptors aligned to the construct of developing clinical sophistication
and independence (Pangaro, 1999; ten Cate, 2006). For example, the anchors in-
cluded: "He/she answers the “What” question and shows basic clinical judgment
following encounter. He/she can be entrusted but only with supervision throughout
the encounter”. The full list of the descriptors is presented in Appendix 2.

<table>
<thead>
<tr>
<th>Anchor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Performed below level expected during Foundation Programme</td>
</tr>
<tr>
<td>2</td>
<td>Performed at the level expected on completion of Foundation Programme / early Core Training PGY-1</td>
</tr>
<tr>
<td>3</td>
<td>Performed at the level expected on completion of Core Training / early higher training PGY-2</td>
</tr>
<tr>
<td>4</td>
<td>Performed at level expected during higher training PGY-3</td>
</tr>
<tr>
<td>5</td>
<td>Performed at level expected on completion of higher training PGY-4</td>
</tr>
</tbody>
</table>

Appendix 2. Construct-aligned scales form (CS)

The setup we used offers two fully crossed (all random) two-facet (assessors and
cases) generalisability designs (A. Alves de Lima, Conde, Costabel, Corso, & van der Vleuten, 2013), one for CS and one for IS. The research protocol was ethically ap-
proved by the Institutional Review Board of our institution.
Analysis

For the ordinal or interval scales (IS) we averaged scores across items of the mini-CEX leading to a case score for each resident. For the construct scales (CS) a case score for each resident was obtained by scores rated by the assessor for each resident. Descriptive statistics were calculated for both scales for each case, assessor, overall case across assessors, and total scores for all cases. Variance component estimations were estimated for each of the seven sources of variance associated with a fully crossed design for both scales (person by case by raters design). For the D-studies (estimating the reliability indices) we used two different designs for both scales: a fully nested design and a design with residents nested within cases and crossed with assessors. In the fully nested design, residents and assessors were nested within cases, because this would enable comparison of our data-set with in vivo data-sets representing different cases (patients) and different assessors (Crossley et al., 2002). In some in vivo conditions, however, there may be only one assessor available for all the residents in the setting, and consequently cases are nested within residents but not within assessors. All analyses were conducted using the mGENOVA software package.

Results

Table 2 shows the means scores and standard errors by year for all cases and all observers.

<table>
<thead>
<tr>
<th>Year</th>
<th>PGY 1</th>
<th>PGY 2</th>
<th>PGY 3</th>
<th>PGY 4</th>
<th>PGY 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Deviation</td>
<td>Mean</td>
<td>Std. Deviation</td>
<td>Mean</td>
</tr>
<tr>
<td>Interval Scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical interviewing</td>
<td>6.22</td>
<td>1.093</td>
<td>6.09</td>
<td>1.104</td>
<td>7.20</td>
</tr>
<tr>
<td>Physical examination</td>
<td>6.78</td>
<td>.667</td>
<td>6.49</td>
<td>1.218</td>
<td>7.60</td>
</tr>
<tr>
<td>Humanistic</td>
<td>7.11</td>
<td>.782</td>
<td>6.49</td>
<td>.944</td>
<td>7.56</td>
</tr>
<tr>
<td>Clinical judgement</td>
<td>6.67</td>
<td>.866</td>
<td>6.62</td>
<td>.984</td>
<td>7.47</td>
</tr>
<tr>
<td>Counselling</td>
<td>6.44</td>
<td>1.014</td>
<td>6.13</td>
<td>1.290</td>
<td>7.51</td>
</tr>
<tr>
<td>Organization</td>
<td>6.33</td>
<td>1.000</td>
<td>6.18</td>
<td>1.134</td>
<td>7.42</td>
</tr>
<tr>
<td>Overall clinical competence</td>
<td>6.44</td>
<td>.726</td>
<td>6.22</td>
<td>.997</td>
<td>7.44</td>
</tr>
<tr>
<td>Constructed- aligned scale scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anchors scores</td>
<td>3.33</td>
<td>.866</td>
<td>3.13</td>
<td>.757</td>
<td>3.84</td>
</tr>
</tbody>
</table>
For both scales, a third of the total variance was associated with universe-score variance, IS 36% vs. CS 29%. The greatest source of variance in the IS was the general error (49%), followed by the main effect of assessors (7%). In the CS, the greatest source of variance was the general error (34%) followed by the assessors’ variability for some residents (23%) (Table 3 and Table 4).

Table 3. Interval scales. Estimated variance components, standard errors, and relative size of variance components

<table>
<thead>
<tr>
<th>Source of variance</th>
<th>Explanation</th>
<th>Estimated variance components</th>
<th>Standard Error</th>
<th>% of total variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vr</td>
<td>Systematic variability of residents</td>
<td>0,51774</td>
<td>0,18354</td>
<td>36</td>
</tr>
<tr>
<td>Vc</td>
<td>Systematic variability of cases (case difficulty)</td>
<td>0,00000</td>
<td>0,01134</td>
<td>0</td>
</tr>
<tr>
<td>Va</td>
<td>Systematic variability of assessors (leniency/stringency)</td>
<td>0,10410</td>
<td>0,09096</td>
<td>7</td>
</tr>
<tr>
<td>Vcr</td>
<td>Variability of residents across cases</td>
<td>0,00000</td>
<td>0,05164</td>
<td>0</td>
</tr>
<tr>
<td>Var</td>
<td>Assessor variability for some residents</td>
<td>0,09167</td>
<td>0,08129</td>
<td>6</td>
</tr>
<tr>
<td>Vca</td>
<td>Assessor variability for some cases</td>
<td>0,02315</td>
<td>0,03352</td>
<td>2</td>
</tr>
<tr>
<td>Vrca</td>
<td>General error term</td>
<td>0,71759</td>
<td>0,11207</td>
<td>49</td>
</tr>
<tr>
<td>Σ</td>
<td></td>
<td>1,45225</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Constructed-aligned scales. Estimated variance components, standard errors, and relative size of variance components

<table>
<thead>
<tr>
<th>Source of variance</th>
<th>Explanation</th>
<th>Estimated variance components</th>
<th>Standard Error</th>
<th>% of total variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vr</td>
<td>Systematic variability of residents</td>
<td>0,24511</td>
<td>0,10594</td>
<td>29</td>
</tr>
<tr>
<td>Vc</td>
<td>Systematic variability of cases (case difficulty)</td>
<td>0,00833</td>
<td>0,01632</td>
<td>1</td>
</tr>
<tr>
<td>Va</td>
<td>Systematic variability of assessors (leniency/stringency)</td>
<td>0,08611</td>
<td>0,07634</td>
<td>10</td>
</tr>
<tr>
<td>Vcr</td>
<td>Variability of residents across cases</td>
<td>0,00489</td>
<td>0,02701</td>
<td>1</td>
</tr>
<tr>
<td>Var</td>
<td>Assessor variability for some residents</td>
<td>0,19167</td>
<td>0,06489</td>
<td>23</td>
</tr>
<tr>
<td>Vca</td>
<td>Assessor variability for some cases</td>
<td>0,02262</td>
<td>0,02120</td>
<td>3</td>
</tr>
<tr>
<td>Vrca</td>
<td>General error term</td>
<td>0,29220</td>
<td>0,04563</td>
<td>34</td>
</tr>
<tr>
<td>Σ</td>
<td></td>
<td>0,85093</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Generalisability coefficients indicated that for both types of scales an approximate sample of seven encounters was needed, assuming both the presence of one different assessor per encounter and the presentation of different cases per encounter (the usual situation in real practice): four encounters when two raters were used and three encounters in case three raters were used (Table 5 and Table 6).
Table 5. Reliability estimates as a function of the number of cases and assessors for the situation where residents are given different cases with different assessors with IS and CS

<table>
<thead>
<tr>
<th>Number of cases</th>
<th>One assessor for all cases</th>
<th>The same two assessors for all cases</th>
<th>The same three assessors for all cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IS</td>
<td>CS</td>
<td>IS</td>
</tr>
<tr>
<td>1</td>
<td>0.37</td>
<td>0.34</td>
<td>0.54</td>
</tr>
<tr>
<td>2</td>
<td>0.54</td>
<td>0.51</td>
<td>0.70</td>
</tr>
<tr>
<td>3</td>
<td>0.64</td>
<td>0.61</td>
<td>0.78</td>
</tr>
<tr>
<td>4</td>
<td>0.70</td>
<td>0.67</td>
<td>0.82</td>
</tr>
<tr>
<td>5</td>
<td>0.75</td>
<td>0.72</td>
<td>0.85</td>
</tr>
<tr>
<td>7</td>
<td>0.80</td>
<td>0.78</td>
<td>0.89</td>
</tr>
<tr>
<td>9</td>
<td>0.84</td>
<td>0.82</td>
<td>0.91</td>
</tr>
<tr>
<td>11</td>
<td>0.87</td>
<td>0.85</td>
<td>0.93</td>
</tr>
<tr>
<td>13</td>
<td>0.88</td>
<td>0.87</td>
<td>0.94</td>
</tr>
<tr>
<td>15</td>
<td>0.90</td>
<td>0.88</td>
<td>0.95</td>
</tr>
</tbody>
</table>

IS: interval scale; CS: Constructed-aligned scale

Table 6. Reliability estimates as a function of the number of cases and assessors for the situation where residents are given different cases but with the same assessors with IS and CS

<table>
<thead>
<tr>
<th>Number of cases</th>
<th>One assessor for all cases</th>
<th>The same two assessors for all cases</th>
<th>The same three assessors for all cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IS</td>
<td>CS</td>
<td>IS</td>
</tr>
<tr>
<td>1</td>
<td>0.39</td>
<td>0.32</td>
<td>0.56</td>
</tr>
<tr>
<td>2</td>
<td>0.53</td>
<td>0.41</td>
<td>0.70</td>
</tr>
<tr>
<td>3</td>
<td>0.61</td>
<td>0.45</td>
<td>0.76</td>
</tr>
<tr>
<td>4</td>
<td>0.66</td>
<td>0.48</td>
<td>0.79</td>
</tr>
<tr>
<td>5</td>
<td>0.69</td>
<td>0.49</td>
<td>0.82</td>
</tr>
<tr>
<td>7</td>
<td>0.73</td>
<td>0.51</td>
<td>0.85</td>
</tr>
<tr>
<td>9</td>
<td>0.76</td>
<td>0.52</td>
<td>0.86</td>
</tr>
<tr>
<td>11</td>
<td>0.77</td>
<td>0.53</td>
<td>0.87</td>
</tr>
<tr>
<td>13</td>
<td>0.79</td>
<td>0.54</td>
<td>0.88</td>
</tr>
<tr>
<td>15</td>
<td>0.80</td>
<td>0.54</td>
<td>0.89</td>
</tr>
</tbody>
</table>

IS: interval scale; CS: Constructed-aligned scale

Discussion

According to the results obtained, and contrary to our expectations, the IS and the CS showed similar performance in terms of sources of variance and in the resulting reliability. Unexplained general error appears to be the major cause of unreliability of both scales, followed by the assessor leniency/stringency in the IS and the assessors' variability for some residents in the CS. The explanation for these results may
be that in our study the assessors were blinded to the training level of the residents. The CS was constructed using predetermined training levels anchors that were accompanied by behavioral descriptors aligned to the construct of developing clinical sophistication and independence. Probably, the knowledge of the level of training of the residents by the assessors could be important in the CS to consolidate a unique frame of reference during observation and rating between assessors to enhance its reliability. Poor inter-rater reliability of clinical skills assessments can be explained if one faculty member rates performance based on PGY level, another uses a standard of self, and another makes a rating based on a gestalt. (Kogan et al., 2011).

There are significant limitations to this study due to specific characteristics of the data-set. Firstly, the small sample size and the resulting precision of variance component estimation (as can be seen from the standard errors in table 3 and 4) diminish the generalisability of the findings to the typical operational application of the mini-CEX. There are additional limitations related to differences between the conditions of the study and those of the mini-CEX in clinical practice: performance was videotaped instead of using direct observation; the information about diagnosis and management plan was obtained in a written format as opposed to face-to-face interview; standardized patients were used; the two-hour training session for the assessors exceeded the usual exposure to such training; the assessors did not know the residents whose performance they judged, while in real practice the assessor-resident relationship tends to inflate scores. Another limitation is that residents belonged to different years of postgraduate medical education and, consequently, differed in expertise, which might have inflated the variance components of residents.

Our study did not replicate the Crossley’s original findings, so it will be important to explore the differences in the contexts and potential limitations on the generalisability of his results.

The phenomenon of construct alignment appears to be less straightforward than might be inferred from Crossley’s initial work. Further exploration of when, where and how it plays a role in improving assessments is be important for future research to establish.
References


Chapter 8
Discussion
Discussion

The studies that together constitute this thesis were performed to shed light on the usefulness of the mini-CEX as a method of workplace-based assessment (WBA). WBA has been defined as multiple, structure and observed assessments, documented immediately following the assessment and performed throughout a clinical rotation. Although this assessment method has a lot of potential for assessing broad range of competences, research is needed to identify formats that offer improved reliability and acceptable feasibility, which would enable broad uses of this assessment method. We specifically investigate: 1) the reliability, 2) the validity, 3) the feasibility and 4) the educational impact of the mini-CEX. In this chapter we return to these four questions, weighing the strength and limitations of the studies and discussing practical implications of the findings for WBA. We also propose direction for future research.

How reliable is the mini-CEX?

The results of our studies indicate that the mini-CEX has reasonable reproducibility with a sample of at least nine encounters, and that a major constraint with regard to its reliability is assessor variance (Chapters 2 and 5). The number of encounters to be assessed is in accordance with the literature, as there is general agreement that between 7 and 15 ratings suffice to achieve a generalizable global estimate of competence, based on a non-systematic sampling of observed assessment. Our laboratory study allowed us to tease out the sources of error and revealed that the assessor was the major source of variation. Understanding the factors impacting on assessors’ judgments and ratings after direct observation is crucial and should be taken into account at the time of implementing the mini-CEX as an assessment tool (Chapter 4). This is in line with Govaerts et al. (2011), who reported that raters make and justify judgments based on personal theories and personal constructs, while their information processing seems to be affected by differences in expertise (Govaerts, Schuwirth, van der Vleuten, & Muijtjens, 2011). Additionally, Gingerich et al. (2011) in a critical review of the literature on rater idiosyncrasy focused on impression formation observed that raters tended to make categorical judgments about ratees as part of impression formation (Gingerich, Regehr, & Eva, 2011). Because of the way raters form impressions, an assessment system requiring ordinal or interval ratings may inadvertently introduce conversion errors due to translation techniques that are unique to individual raters. Along the same lines, Crossley et al. (2011) demonstrated that in real life settings, mini-CEX scales constructed in accordance with categorical judgments which reflect the development of clinical sophistication and independence, have higher reliability than ordinal or interval.
ratings (Crossley, Johnson, Booth, & Wade, 2011). The construct of clinical sophistication emphasizes progressive expectations as a student or a resident progresses towards eventual independence through the clinical years and through the residency. In relation with the construct of independence or entrustability, ten Cate (2006) argues that clinical supervisors’ judgments focus on the concept of entrustability. Entrusting a critical activity should lead to the trainee being granted responsibility in all similar future circumstances. Once sound feedback has confirmed a critical number of times that all went well, the entrustment could be formalised and considered a qualification to act independently (ten Cate, 2006).

We tested this hypothesis with the aim of enhancing reliability of assessment, but did not find the expected higher reliability due to the use of nominal scales (Chapter 7). These results may be attributable to methodological issues. In our study, the assessors were blinded to the level of expertise of the residents, and assessors’ knowledge of residents’ levels of expertise may play an important role as a frame of reference in enhancing reliability. These findings reinforce the prominence of the role of the rater as a source of unreliability in the mini-CEX. The value of assessment appears to be determined by the users of the instrument rather than by the instrument. Unsystematic observation and documentation of student and resident performances, poor rater consistency, homogeneity of the population of interest, small sample sizes, and factors related to the test occasion and the timeliness of testing were all found to negatively affect the reliability of the mini-CEX. On a more philosophical level, the question might be posed if it is possible and defensible for assessors to be standardized, trained, and benchmarked. In psychometric models there is only one truth, called the ‘true score’, and any deviation from that score is dismissed as ‘error’. However, as Govaerts has made very clear, assessors are not passive ‘instruments’, but rather ‘creators’ of their own judgments. This discussion is similar to the development of ideas around learning theories, which have moved from classic mastery-oriented learning to constructivist learning: learners are the creators of their own idiosyncratic knowledge and expertise. Constructivist learning is probably even more relevant to learning in the workplace, where learning is grounded in learners’ authentic experiences and reflections on those experiences. So what is considered noise or error in psychometric theory may be the natural result of idiosyncratic experiences and created realities not only for the learner but also for the assessor.

**Practical recommendations**

The outcomes of our reliability studies suggest that around ten encounters suffice for a reproducible outcome. In terms of testing time, ten encounters compare favourably with the samples needed for other standardized and objectified assess-
ment formats, although one would expect poor reliability of an instrument that, like the mini-CEX, is characterized by absence of explicit characteristics. Assessor training may have some advantages. Firstly, even though the literature offers no hard evidence to this effect, it is generally assumed that it can improve inter-rater agreement, which in turn improves reliability of the format. Secondly, training may have a beneficial effect on the feasibility of assessment, because training sessions offer an opportunity to extensively introduce, explain, and discuss the method with designated assessors, thereby possibly enhancing examiners’ commitment to the application of the format.

**Directions for future research**

In future studies, we should rethink the value of psychometric theory, particularly in relation to work-based assessment, a suggestion that has also been made by others (Govaerts, van der Vleuten, Schuwirth, & Muijtjens, 2007; Schuwirth & van der Vleuten, 2006). The results of our studies indicate that assessor’s variance is a major constraint with regard to the reliability of the mini-CEX. In this sense, it would be interesting to better understand what influences observer’s judgments behaviours after direct observation, what type of observers we should identify and what type of observers’ training would be necessary to establish reliable measurements of clinical performance.

**How valid is the mini-CEX?**

In chapter 2, we have demonstrated evidence for the construct validity of the mini-CEX when used across the years of a residency program. The results showed modest increases of the mean scores on the mini-CEX across years of training. This is consistent with findings from other authors as Kogan et al. (2003), who reported correlation coefficients ($r = 0.17-0.43$) for medical students’ mini-CEX scores and scores on written and clinical performance exams (Kogan, Bellini, & Shea, 2003). In addition, Hatala et al. (2006) computed correlation coefficients from 0.29 to 0.60 for residents’ mini-CEX scores and their scores in internal medicine oral, bedside, and written exams (Hatala, Ainslie, Kassen, Mackie, & Roberts, 2006). Al Ansari et al. (2013) conducted a meta-analysis of published studies to determine the construct and criterion validity of the mini-CEX (Al Ansari, Kauser & Donnon, 2013), and found that the mini-CEX was a useful in-training assessment tool with clear evidence of construct and criterion related validity.
Practical recommendations

Practical recommendations to enhance the validity of the mini-CEX include assessment of students or residents across multiple encounters and different conditions (existing, new, chronic or acute) across age groups and gender, with problems that differ in levels of complexity, across systems/disciplines, and across clinical settings (office, ER or inpatient, among others).

Directions for future research

We think that the validity of the mini-CEX needs further exploration. It is necessary to determine the appropriate number of encounters to observe and the number of patients with different conditions (existing, new, chronic, or acute), across age groups and gender, with problems that differ in levels of complexity and across systems/disciplines, or across clinical settings (office, ER or inpatient, among others) to sample in order to enhance the validity of the mini-CEX. Further research on the efficacy of clearly training examiners is needed; research should also focus on the cultural and environmental influences on rating and on the potential value of quality assurance and feedback process for examiners.

How feasible is the mini-CEX?

Regarding feasibility (Chapter 2), our results showed that it was not possible to achieve the number of encounters required. One of the explanations for this may be that the mini-CEX was completely new to the mostly patient-care orientated institutions where it was introduced in our study. We developed only written instructions and never discussed with the teachers who were to act as observers the different competencies they were expected to focus on. Prior to the implementation of the assessment method future examiners were not asked for their opinions about the feasibility of the format. The residency programs were not scrutinized in all the hospitals, and the examiners themselves were expected to arrange for scheduling of the mini-CEX encounters in their daily activities. Feasibility studies reported in the literature focus mostly on completion rates of the instrument or user satisfaction, and 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, reported completion rates of 96.4% and 100%, respectively, but Turnbull et al. (2000) concluded that feasibility was good with a response rate of only 23% (Durning, Cation, Markert, & Pangaro, 2002; Torre, Simpson, Elnicki, Sebastian,
Holmboe, 2007; Turnbull & Van Barneveld, 2002). Wilkinson et al. (2008) attributed feasibility problems to lack of time and the fact that the procedure was experienced as time consuming. They concluded that assessment instruments like the mini-CEX must be well integrated within the curriculum and embedded in practice routines (Wilkinson et al., 2008).

**Practical recommendations**

Greater involvement in the implementation process of assessors, members of the staff and perhaps senior residents with experience in assessment in particular may be advisable. In-depth interviews with staff members and senior residents might have provided them with detailed information about the competencies to be assessed, identified aspects that were important for them to observe, and helped to reach agreement on minimum requirements for residents according to their level of expertise (Chapter 3). Furthermore, repeating in-depth interviews during the implementation of the method might have resulted in early detection of problems and thus allowed for early intervention to resolve them. In addition, workshops would have been a better way to introduce the implementation of an instrument than written instructions. In summary, the feasibility of a new performance-based assessment format seems likely to benefit from taking account of future examiners’ opinions about the feasibility of the method and about the potential interference of assessment with their daily activities. The take-home message seems to be that it is important to intensify examiner involvement in the introduction of a new method and to pilot a method prior to its definitive implementation.

**Directions for future research**

Several questions remain for a future research agenda on feasibility issues: which are the limits to feasibility; which models would enhance the environment for assessment and the quality of the information obtained?

**Which is the educational impact of the mini-CEX?**

Assuming that assessment is a potentially powerful driver of students’ or residents’ educational behaviours, we expected the introduction of an instrument for workplace-based assessment, like the mini-CEX, to show an impact on learning styles and clinical skills. In terms of learning styles, we found positive results among residents, with exposure to the mini-CEX stimulating residents to try and construct
their own overall picture of the topics and to develop skills for self-directed learning (Chapter 3). However, in terms of clinical behaviour, we found no statistically significant effects in any of the domains (Chapter 6). A possible explanation may be differences between the curriculum in action and the intended curriculum, with lack of awareness of the program in action which perhaps might have played a crucial role in bringing about the absence of effects of the assessment tool. Another explanation may be that behavioural changes are most likely to occur when many factors are explicitly targeted. Individual factors and the quality and the context of the feedback can have a profound effect on the magnitude of the response. In chapter 6 the observers were assumed to provide appropriate feedback, but although all of them had documented experience with the mini-CEX, their levels of clinical competence differed. We do not know exactly how much supervision and feedback residents or students need to attain an adequate level of competence, but it seems that the low frequencies that were applied in chapter 6 may have been inadequate to achieve measurable behavioural changes. Our conclusions are in accordance with those of other authors. In a systematic review, Miller et al. (2010) found that studies examining the mini-CEX showed largely positive results in terms of learner satisfaction but failed to demonstrated measurable changes in attitudes, skills, knowledge, or behaviour (Miller & Archer, 2010). Previous reviews investigating tools for direct observation and assessment of clinical skills also reported that few studies reported educational outcomes (Kogan, Holmboe, & Hauer, 2009; Pelgrim et al., 2011).

**Practical recommendations**

This thesis shows that even when measures taken to improve assessment during residency or clerkships are effective, it should not be taken for granted that the assessment tool will actually achieve the desired effects on clinical behaviour. High quality feedback requires sufficient medical competence to provide reflection on performance and, for beginners, direction on performance (Pelgrim, Kramer, Mokkink, & van der Vleuten, 2012a, 2012b). Furthermore, teaching skills such as setting learning objectives and providing structure for supervision and feedback are necessary to create a positive environment for feedback (Kilminster, Jolly, & van der Vleuten, 2002). Observation should be followed by clear and timely feedback combined with opportunities for learners to practice the competence and demonstrate at a later date their progress on weaknesses identified by assessment. Additional measures aimed at supporting changes in these aspects, such as identifying what is important to assess, improved agreement on minimum student requirements, and feedback training for observers might have facilitated behavioural change. If such training had been provided, the assessment might have triggered new behaviours,
and this might have been facilitated by extensive training sessions for prospective observers.

**Directions for future research**

Serious consideration should be given to the use of study designs that are able to demonstrate conclusively that workplace-based assessment methods like the mini-CEX are associated with performance improvement. Such studies should be carefully designed with appropriate size to be able to show significant changes over extended periods and should involve exposure of matched groups of learners to different interventions. This may be achievable when faculty development is supported and hospitals and medical societies accept that it is one of their main tasks to encourage discussions on the position of medical education.

**Final conclusion**

The results of our studies have shown that the mini-CEX can evaluate a wide range of learners’ competencies in a wide variety of clinical settings and with a diverse set of problems in a manner that is reliable and valid. Multiple encounters with different examiners and patients have been found to produce reliable data. A good understanding of the factors impacting on assessors’ judgments and ratings after direct observation is crucial and should be taken into account by those developing and organizing assessor training sessions. Close monitoring of the assessment program in action and supporting and facilitating intended changes, e.g., by supervisor training and encouraging discussions of the position of medical training as one of the main tasks in hospitals, are important measures that may enhance the effects of assessment on the learning environment in clerkships and residency programs. Attention should be given to all aspects of the assessment environment in which the assessment tool is implemented and to identify measures that can support the intended effects. Implementing educational change in residency or clerkships is a difficult task. In the workplace, assessors, residents and students often feel forced to adopt new educational behaviours not because they want to, but because some educationalist has come up with an innovative idea. For successful implementation of educational change we should adopt an active participation of educationalists, doctors, residents, and students who are explicitly involved, motivated to take part in the introduction of the new educational format and in the evaluation of the effects of innovations.

Medical educators find themselves faced with the challenge of using in-training assessment to evaluate medical students’ and residents’ clinical skills based on
their performance in patient encounters. We think that the mini-CEX should never be used on its own as a separate assessment tool but should always be combined with other exercises that trainees may perform after patient encounters, such as oral case presentations, written exercises that assess clinical reasoning, and literature searches. In addition, video review of patient encounters appears to offer a powerful means of evaluating and providing feedback on trainees’ skills in clinical encounters. Such an approach may improve reliability of assessment by balancing the weights of different assessment formats. Finally, a continued positive impact of workplace-based assessment like the mini-CEX can only be ensured when measures are taken to incorporate assessments in the clinical routines of clinicians and residents.
References


Summary
The subject of this thesis is the assessment of clinical competence based on direct observation and feedback of a learner interacting with a patient in undergraduate and postgraduate medical training. The research focuses on the mini-CEX because this instrument is designed around the skills most commonly performed by residents and students in actual patient encounters and the educational interactions that attending physicians routinely have with residents or students during teaching rounds. This thesis focuses on the validity, reliability, feasibility, and educational impact of this assessment tool.

**Chapter 1** presents a historical overview of assessment, a definition of clinical competence, and a detailed description of the mini-CEX as a work-based assessment tool. Performance-based assessments are the main assessment tools used in residency programs because the context of these programs offers excellent opportunities for assessment in authentic environments. Residents and students spend a major portion of their medical training working and learning with patients during residencies and clerkships, which make a central contribution to the development of clinical competence.

However supervision and feedback are infrequent and deficient in structure and continuity. The strength of existing assessment tools for in-training evaluation, including the mini-CEX, is that they are inexpensive and easy to use, although they require some developmental resources and a fair amount of raters’ time for training and implementation. The mini-CEX provides an opportunity to evaluate continuous practice performance based on the assumption that the skills to be demonstrated are very close to those required for clinical practice. The areas of weakness include the psychometric properties of this tool, namely reliability and validity. Given the challenges associated with current in-training evaluation practices, the meaningfulness of work-based assessment tools like the mini-CEX as instruments for summative or formative assessment should be reconsidered. We address the following specific research questions:

1. How reliable is the mini-CEX?
2. How valid is the mini-CEX?
3. How feasible is the mini-CEX?
4. What is the educational impact of the mini-CEX?

**Chapter 2** addresses the validity, reliability, feasibility, and satisfaction rates of the mini-CEX in a cohort of 108 cardiology residents. The residents were from seventeen cardiology residency programs in Buenos Aires. Validity was evaluated by determining the capability of the instrument to discriminate between pre-existing levels of clinical seniority. Generalizability theory was used to determine reliability. Feasibility was defined by a minimum number of completed observations: 50% of residents obtaining at least four mini-CEXs. Evaluators and residents rated their
satisfaction with the instrument on a one to nine rating scale. The total number of encounters was 253. We found that the mini-CEX discriminated between pre-existing levels of seniority, required considerable sampling to achieve sufficient reliability, and was not feasible within the current circumstances, although it was considered a valuable assessment tool as indicated by the evaluators’ and residents’ satisfaction ratings.

Chapter 3 describes a qualitative exploration of residents’ perceptions of the mini-CEX as an assessment tool and its influence on their approach to learning and studying. All sixteen residents from a cardiology training program were included. The results show that, in all cases, residents demonstrated an intrinsic interest in the subject matter. They showed self-regulating strategies when required to select, relate and make critical appraisals of their own. They consistently demonstrated an aim to build a relationship between individual experience and their chosen topic. The residents felt comfortable using the mini-CEX because it melded with their routine. They found the mini-CEX to be a useful assessment tool with a favorable influence toward a constructive approach to learning.

Chapter 4 describes a qualitative exploration of teachers’ perceptions of the value of the mini-CEX as an assessment tool, its feasibility, and the influence of the format on teachers’ pass/fail decisions and feedback delivery. Seventeen teachers who had all acted as evaluators at least twice during the study period were interviewed. The study shows that teachers perceived the mini-CEX as a feasible and useful formative assessment tool. Contextual factors such as interpersonal bonds, preconceived performance notions, and lack of specific guidelines and performance standards provide insight into why teachers tended to be reluctant to fail poorly performing residents.

In Chapter 5, reliability estimations of the mini-CEX in a controlled setup are investigated. In real life settings, reliability estimations are based on the assumption of local independence: the object of the measurement should not be influenced by the measurement itself and samples should be completely independent. This is, however, difficult to achieve. Furthermore, the variance due to case/patient or assessors is completely confounded. We have no idea of the extent to which each of these factors contributes to the noise in the measurement. This study used a controlled setup to overcome these difficulties in estimating the reproducibility of the mini-CEX. Twenty-one residents were videotaped during the same three encounters with three different patients. Three assessors assessed all the encounters. This yielded a fully crossed (all random) two-facet generalizability design. A quarter of the overall variance was associated with universe score variance (28%). The largest source of variance was the general error term (34%), followed by the main effect of assessors.
Generalizability coefficients indicated that a sample of approximately nine encounters was needed with one single assessor per encounter and different assessors and different cases for each encounter, as is common in real practice; four encounters were needed with two raters, and three encounters with three raters. Unexplained general error and the leniency/stringency of assessors were the major causes of unreliability in the mini-CEX.

Chapter 6 addresses changes in clinical skills proficiency in a cohort of fourth-year medical students as assessed with the mini-CEX during clerkship. The students were divided in two consecutive groups. Both groups received the same contents and teaching strategies, but group 1 was and group was not assessed with three formative mini-CEXs during the clerkship period. At the end of the three weeks each group was summatively assessed using a mini-CEX with a simulated patient observed by two assessors. We found no significant differences in clinical performance and behavior between the two groups. The low number of encounters and the short exposure to the mini-CEX may have influenced the results.

Chapter 7 addresses the reliability of the mini-CEX using interval and construct-aligned scales. Three encounters were videotaped from 21 residents. Each encounter was assessed by 3 assessors who assessed all encounters for all residents twice. The first time using the interval scale and 30 days later with the construct-aligned scale. Each assessor was blinded to the training level of the residents. This delivered a fully crossed (all random) two-facet generalizability design each time. The universe score variance was 36% for interval scale and 29% for construct-aligned scale. In the interval scale the general error was 49% followed by assessor variance (7%). In the construct-aligned scale the general error was 34% followed by the assessors’ variability for some residents (23%). Generalizability coefficients indicated that for both types of scales a sample of 7 encounters was needed, assuming both the presence of one different assessor per encounter. Both scales showed similar performance in terms of sources of variance and in the resulting reliability. The explanation for these results may be that assessors were blinded to the level of training of the residents. The knowledge of the level of training of the residents by the assessors could be important in the construct-aligned scale to consolidate a unique frame of reference during observation and rating between assessors to enhance its reliability.

In Chapter 8, the findings of this thesis are discussed and recommendations are made for future research. The mini-CEX can evaluate a wide range of clinical competencies in a wide variety of clinical settings and with a diverse set of problems in a manner that is reliable and valid. Multiple encounters with different examiners and patients have been found to produce reliable data. A good understanding of the
factors impacting on assessors' judgments and ratings after direct observation is crucial and should be taken into account in organizing assessor training sessions. Close monitoring of the assessment program-in-action and supporting and facilitating intended changes - for example by training supervisors and encouraging discussions of the position of medical training as one of the main tasks of hospitals - are important measures that might improve the effects of assessment on the learning environment in clerkships and residency programs. Attention should be given to all the aspects of the assessment environment in which the assessment tool is implemented to identify measures that can support and enhance the intended effects.
Samenvatting
Het onderwerp van dit proefschrift is toetsing van klinische competentie in de werkplaats met directe observatie en feedback tijdens de basisarts- en vervolgopleiding. De onderzoeken die besproken worden, hebben alle betrekking op de Korte Klinische Beoordeling (KKB), omdat het ontwerp daarvan aansluit bij de vaardigheden die AIOS en studenten normaal uitvoeren tijdens patiëntencontacten en die ook aan de orde komen tijdens het onderwijs door begeleiders en stafleden in de kliniek. De nadruk in de onderzoeken ligt op de validiteit, betrouwbaarheid, uitvoerbaarheid en onderwijskundige effecten van dit beoordelingsinstrument.

In **Hoofdstuk 1** wordt een historisch overzicht gegeven waarin aandacht besteed wordt aan het beoordelen en definiëren van klinische competentie en een gedetailleerde beschrijving gegeven wordt van de KKB als instrument voor werkplaatsbeoordeling. Toetsing in de werkplaats is de belangrijkste toetsvorm tijdens de vervolgopleiding omdat de leeromgeving zich hier bij uitstek voor leent. Voor zowel AIOS als studenten vormen patiëntencontacten in de klinische praktijk de kern van de leeromgeving waardoor deze contacten de belangrijkste bijdrage leveren aan de ontwikkeling van de klinische competentie van toekomstige artsen en specialisten.

Helaas laten in het klinisch onderwijs de frequentie, structuur en continuïteit van supervisie en feedback nogal wat te wensen over. Daartegenover staan sterke punten van de bestaande instrumenten voor klinische toetsing, waaronder de KBB, zoals de geringe kosten en het gebruiksgemak, al dienen er wel middelen beschikbaar gesteld te worden voor training van beoordelaars en een zorgvuldige implementatie van instrumenten. De KBB maakt longitudinale toetsing van het functioneren van AIOS en studenten mogelijk op voorwaarde dat KBBs vaardigheden bevatten die overeenkomen met en relevant zijn voor de klinische praktijk. De zwakke punten van de KBB zijn de psychometrische eigenschappen, zoals betrouwbaarheid en validiteit. Gegeven de eisen die in de huidige medische opleidingen aan toetsing gesteld worden, verdient het aanbeveling om instrumenten voor werkplaatsbeoordeling, zoals de KBB, aan een grondige evaluatie te onderwerpen. In dit proefschrift is daarom een antwoord gezocht op de volgende onderzoeksvragen:

1. Hoe betrouwbaar is de mini-CEX?
2. Hoe valide is de KBB?
3. Hoe praktisch uitvoerbaar is de KBB?
4. Wat zijn de onderwijskundige effecten van de KBB?

**Hoofdstuk 2** beschrijft een onderzoek naar de validiteit, betrouwbaarheid en uitvoerbaarheid van de KBB en de tevredenheid van AIOS met dit instrument. Aan dit onderzoek werd deelgenomen door een cohort van 108 AIOS uit zeventien cardiologieopleidingen in Buenos Aires. Om de validiteit te beoordelen werd gemeten in hoeverre het instrument onderscheid maakte tussen de competentieniveaus van
AIOS in verschillende fasen van de opleiding. De betrouwbaarheid werd bepaald met behulp van generaliseerbaarheidstheorie. De uitvoerbaarheid werd beoordeeld door na te gaan of een minimaal aantal beoordelingen plaatsvond: ten minste vier KBBs voor 50% van de AIOS. De beoordelaars en de AIOS scoorden hun tevredenheid met het instrument op een negenpuntsschaal. Tijdens het onderzoek werd een totaal aantal van 253 KBBs uitgevoerd. Het bleek dat de KKB onderscheid maakte tussen bestaande competentieniveaus, een aanzienlijk aantal beoordelingen vereiste om voldoende betrouwbaarheid te bereiken en niet goed uitvoerbaar was onder de bestaande condities, al gaven in het tevredenheidsonderzoek zowel beoordelaars als AIOS aan dat vanuit hun optiek de KBB een waardevol beoordelingsinstrument was.

**Hoofdstuk 3** beschrijft een kwalitatief onderzoek waarin van zestien AIOS die de cardiologieopleiding in hetzelfde ziekenhuis volgden de mening werd gepeild over de KKB als beoordelingsinstrument en de leereffecten van de KKB. De AIOS toonden intrinsieke belangstelling voor de onderwerpen die aan de orde kwamen. Ze maakten gebruik van strategieën voor zelfregulering bij opdrachten waarbij zij moesten selecteren, verbanden leggen en kritisch oordelen. Een systematische bevinding was dat de AIOS doelbewust verbanden probeerden te leggen tussen hun persoonlijke ervaringen en het onderwerp dat zij gekozen hadden. De AIOS vonden de KKB goed bruikbaar omdat deze naadloos ingepast kon worden in de werkomgeving. Ze waren van oordeel dat de KKB een nuttig beoordelingsinstrument was met een gunstig effect op een constructieve benadering van leren.

**Hoofdstuk 4** beschrijft een kwalitatief onderzoek waarin docenten gevraagd werd hun mening te geven over de KKB: de waarde als beoordelingsinstrument, de toepasbaarheid en de invloed op zak/slaagbeslissingen en het geven van feedback. Zeventien klinisch docenten werden geïnterviewd die minimaal twee keer tijdens de onderzoeksperiode optraden als beoordelaars van een KKB. De docenten waren van oordeel dat de KKB een bruikbaar en nuttig instrument was voor formatieve beoordeling. Omgevingsfactoren zoals persoonlijke verhoudingen, vooropgezette ideeën over prestatieniveaus en het onthrekken van duidelijke richtlijnen en beoordelingsstandaarden verklaarden de sterke geneigdheid van docenten om AIOS met slechte prestaties toch niet te laten zakken.

Het onderzoek in **Hoofdstuk 5** betrof een meting van de betrouwbaarheid van de KKB in een experimentele opzet. In realistische situaties gaan bepalingen van betrouwbaarheid uit van plaatselijke onafhankelijkheid: het meetobject wordt niet beïnvloed door het meetinstrument en steekproeven zijn volkomen onafhankelijk, maar deze situatie is in de praktijk niet goed haalbaar. Bovendien is niet duidelijk welk deel van de variantie gerelateerd is aan casus/patiënten of beoordelaars, om-
dat onbekend is wat de omvang is van de bijdrage van deze factoren aan de ruis in de meting. Om deze problemen te omzeilen werd een experimenteel onderzoek uitgevoerd om de reproduceerbaarheid van de KKB te meten. Van elk van 21 AIOS werden video-opnames gemaakt tijdens dezelfde drie contacten met dezelfde drie verschillende patiënten. Alle contacten werden door dezelfde drie beoordelaars beoordeeld. Dit resulteerde in een volledig gekruist (geheel gerandomiseerd) generaliseerbaarheidsmodel met twee facetten (beoordelaars en casus). Een kwart van de variantie werd verklaard door de variantie in de ‘universe’-score (28%). De grootste bron van variantie was de algemene foutenterm (34%) gevolgd door het effect van de beoordelaars (streng/mild) (18%). Op basis van de generaliseerbaarheidscoëfficiënten kan geconcludeerd worden dat een steekproef van ongeveer negen casus voldoende is bij één beoordelaar per contact en steeds verschillende casus (de normale situatie in de praktijk); vier contacten zijn voldoende bij twee verschillende beoordelaars en drie contacten bij drie verschillende beoordelaars. De onverklaarde algemene foutenterm en de strengheid/mildheid van de beoordelaars vormden de belangrijkste bronnen voor onbetrouwbaarheid van de KKB.

**Hoofdstuk 6** beschrijft een onderzoek naar het optreden van veranderingen in de beheersing van klinische vaardigheden in een cohort vierdejaarsstudenten na beoordeling door middel van KKBs. De studenten werden verdeeld in twee opeenvolgende groepen die beide wat inhoud en vorm betreft hetzelfde onderwijs (coassistentschap) volgden. De eerste groep werd in de loop van de drie weken van het coassistentschap beoordeeld door middel van drie formatieve KKBs. De tweede groep nam niet deel aan deze formatieve KKBs, maar uitsluitend, samen met de eerste groep, aan de summative KKB aan het eind van het coassistentschap. De summative KKB behelsde een contact met een simulatiepatiënt, dat beoordeeld werd door twee beoordelaars. Er werden geen significante verschillen gevonden tussen de twee groepen in klinische competentie en gedrag zoals beoordeeld met de summative KKB. Dit resultaat is wellicht te verklaren door het geringe aantal patiëntencontacten en de beperkte duur van de onderzoeksperiode.

**Hoofdstuk 7** betreft een onderzoek naar de betrouwbaarheid van KKBs met behulp van intervalbeoordelingsschalen en constructspecifieke schalen. Van elk van 21 AIOS werden drie patiëntencontacten op video opgenomen en beoordeeld door drie beoordelaars die alle contacten van alle AIOS tweemaal beoordeelden. De eerste keer werd gescoord op een intervalschaal en de tweede keer, dertig dagen later, op een constructspecifieke schaal. Elke beoordelaar was niet op de hoogte van het trainingsniveau van de AIOS. Dit resulteerde in een volledig gekruist (volledig gerandomiseerd) generaliseerbaarheidsmodel met twee facetten. De variantie in de ‘universe’-score was 36% bij de intervalschaal en 29% bij de constructspecifieke schaal. De variantie in de algemene foutenterm verklaarde bij de intervalschaal
49% van de variantie en de beoordelaarsvariantie verklaarde 7%. Bij de construct-specifieke schaal waren deze percentages respectievelijk 34% en 23%. De generaliseerbaarheidscoëfficiënten gaven aan dat zeven KBBs voldoende waren voor een betrouwbaar oordeel met één beoordelaar per contact. Beide schalen gaven vergelijkbare resultaten betreffende variantiebronnen en betrouwbaarheid. Een mogelijke verklaring hiervoor is dat de beoordelaars niet op de hoogte waren van het trainingsniveau van de AIOS. Deze kennis zou belangrijk kunnen zijn om bij de constructspecifieke schaal de beoordelaars een homogeen referentiekader te bieden voor observatie en beoordeling en zo de betrouwbaarheid van de KBB te verhogen.

In Hoofdstuk 8 worden de resultaten van de onderzoeken besproken en aanbevelingen gedaan voor verder onderzoek. Uit de resultaten komt naar voren dat de KBB met een grote variatie aan casus en in zeer verschillende klinische omgevingen geschikt is voor betrouwbare en valide beoordeling van de competentie van aankomend artsen en specialisten. Betrouwbare beoordeling vereisen een aantal verschillende beoordelaars en verschillende casus. Een goed begrip van de factoren die van invloed zijn op het beoordelen en scoren is van groot belang en het verdient aanbeveling dergelijk inzicht als uitgangspunt te nemen bij het ontwikkelen en organiseren van trainingen voor beoordelaars. Reguliere evaluatie van toetsprogramma’s in de praktijk en ondersteuning en begeleiding van beoogde veranderingen - bijvoorbeeld door middel van docententrainingen en het stimuleren van discussies over de status van het medisch onderwijs als een van de hoofdtaken van ziekenhuizen - zijn belangrijke voorwaarden om te bereiken dat toetsing een gunstig effect heeft op de leeromgeving van coassistentschappen en specialistenopleidingen. Het verdient aanbeveling om aandacht te besteden aan alle aspecten van de leeromgeving waarin een toetsinstrument wordt toegepast om goed te kunnen beoordelen welke maatregelen genomen kunnen worden om de beoogde effecten zinvol te ondersteunen.
Curriculum Vitae
Alberto Alves de Lima was born on November 26, 1963, in Buenos Aires, Argentina.

He obtained his secondary school diploma in 1981 at the San Juan el Precursor School in San Isidro, Buenos Aires. In 1983 he started studying Medicine at the University of Buenos Aires. In 1988 he graduated as a medical doctor and in 1994 he was board certified in cardiology. In 1995 concurrently with his patient care job he started working in the postgraduate educational department of the Instituto Cardiovascular de Buenos Aires. In 2002 he earned a Master’s Degree in Health Professions Education from Maastricht University, and in 2003, he started his PhD thesis at the Instituto Cardiovascular de Buenos Aires under the supervision of Professor Cees van der Vleuten.

He is currently vice-chairman of the Cardiology Department and Director of the Educational and Research Department of the Instituto Cardiovascular de Buenos Aires. Alberto is married to Ingrid Gerold and has two children, Tomás aged 19 and Agustina aged 15.