Adv in Health Sci Educ (2013) 18:5–13
DOI 10.1007/s10459-011-9343-y

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

Alberto Alves de Lima · Diego Conde · Juan Costabel · Juan Corso · Cees Van der Vleuten

Received: 8 December 2011 / Accepted: 8 December 2011 / Published online: 23 December 2011

Abstract 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.

Keywords Reliability · Mini-CEX · Full crossed design · Laboratory study · Residents · Post-graduate training

A. Alves de Lima (✉) · D. Conde · J. Costabel · J. Corso
Instituto Cardiovascular de Buenos Aires, Blanco Encalada 1525, 1428 Ciudad de Buenos Aires, Buenos Aires, Argentina
e-mail: aealvesdelima@fibertel.com.ar

C. Van der Vleuten
Maastricht University, Maastricht, The Netherlands

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 work-based assessment, including the mini-CEX, Pelgrim et al. (2011) found eight studies reporting reliability results, showing mostly acceptable (>0.8) reliability with a feasible sample size of ten encounters.

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 et al. (2010), who revealed a reproducibility coefficient of 0.23 (0.70 for 10 assessors or encounters). Hill et al (2009) analyzed a total of 3,400 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. Weller et al. (2009) 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. Moreover, Kogan et al. (2009), 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). The only study to examine reliability in a controlled laboratory setting (Margolis et al. 2006) reported less favorable results from an analysis of ratings of a total of 48 cases by eight practicing physicians. 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 Sates 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 3 h. 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.

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 et al. (2006) 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 4-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 7 days after an uncomplicated acute myocardial infarction, a 37-year-old dyslipedemic female attending the clinic for a blood pressure check-up and for blood results 1 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 2 h. 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 et al. 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.

Discussion

We examined assessment data in a fully crossed design in which every resident was assessed by the same three assessors to asses 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 assessees) 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

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

		Medical interviewing skills	Physical examination skills	Humanistic qualities/ professionalism	Clinical judgement	Counselling skills	Organization/ efficiency	Global clinical competence
Case 1								
A 1	M	6.61	6.9	7.61	6.09	6.14	6.14	6.04
	SD	1.85	1.25	0.58	2.09	2.05	1.79	1.68
A 2	M	6.14	6	6.19	6.3	6.38	6.23	6.57
	SD	1.45	0.5	1.5	1.34	1.85	1.6	1.59
A 3	M	5.95	6.11	6.04	2.96	5.71	6	5.42
	SD	1.85	1.28	1.8	1.72	2.14	1.78	1.8
O	M	6.24	6.25	6.62	6.06	6.08	6.13	6.02
	SD	1.79	1.22	1.54	1.73	2.01	1.7	1.73
Case 2								
A 1	M	7.33	7	7.66	6.52	6.9	6.57	6.52
	SD	1.15	1.22	0.57	2.33	1.17	1.2	1.32
A 2	M	6.09	5.95	5.71	5.8	5.9	5.95	6.19
	SD	1.13	0.38	1	0.81	1.41	0.97	1.12
A 3	M	4.9	6.23	5.76	5.04	5	5.66	4.8
	SD	1.48	1.3	1.3	1.43	1.3	1.31	1.24
O	M	6.11	6.4	6.38	5.79	5.81	6.06	5.84
	SD	1.59	1.12	1.34	1.73	1.65	1.21	1.42
Case 3								
A 1	M	6.95	7.47	7.61	6.42	6.09	6.42	6.61
	SD	1.96	0.87	0.49	1.85	2.34	1.69	1.62
A 2	M	5.71	6	5.71	5.9	5.33	5.76	5.71
	SD	0.84	0	0.9	0.53	1.35	0.53	0.78

Table 1 continued

		Medical interviewing skills	Physical examination skills	Humanistic qualities/ professionalism	Clinical judgement	Counselling skills	Organization/ efficiency	Global clinical competence
A 3	M	5.52	5.61	6.14	5.38	4.71	5.23	5.04
	SD	1.36	0.97	1.15	1.49	1.84	1.67	1.65
O	M	6.06	6.37	6.49	5.81	5.38	5.81	5.79
	SD	1.57	1.09	1.2	1.63	1.94	1.46	1.53
All cases								
Total score	M	6.14	6.34	6.5	5.92	5.8	6	5.88
	SD	1.62	1.14	1.37	1.64	1.84	1.47	1.56

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

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

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

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

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

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

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. Similar assessor effects were found by Weller et al. (2009), based on analysis of data from 331 assessments forms, 61 trainees, and 58 assessors, revealing variance of assessor stringency to be the main cause of unreliability, contributing 40% to score variation. In an analysis of a total of 3,499 mini-CEX forms, Hill et al. (2009) found a considerable contribution (29%) of assessor stringency to the score variation as well, from which they inferred some practical implications. 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 minimum 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 (Hill et al. 2009) 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.

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) in 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 2-h 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.

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 et al. on the fact that assessors training might in some way be helpful towards optimizing reliability. 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 et al. (2011) 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.

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

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

Cook, D. A., Beckman, T. J., Mandrekar, J. N., & Pankratz, V. S. (2010). Internal structure of the mini-CEX for internal medicine residents: factor analysis and generalisability. *Advances in Health Sciences Education: Theory Practice, 15*, 633–645.

Cook, K. F., Kallen, M. K., & Amtmann, D. (2009). Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality. *Assumption Quality of Life Research, 18*, 447–460.

Crossley, J., Davies, H., Humphris, G., & Jolly, B. (2002). Generalisability: a key to unlock professional assessment. *Medical Education, 36*, 972–978.

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

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

Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: a conceptual model. *Medical Education, 45*, 1048–1060.

Kogan, J. R., Holmboe, E. S., & Hauer, K. (2009). Tools for direct observation and assessment of clinical skills of medical trainees. *Journal of the American Medical Association, 302*, 1316–1326.

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

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

Pelgrim, E. A. M., Kramer, A. W. M., Mokkinik, H. G. A., Van den Elsen, L., Grol, R. P. T. M., & Van der Vleuten, C. P. M. (2011). In-training assessment using direct observation of single-patient encounters: A literature review. *Advances in Health Science Education: Theory and Practice, 16*, 131–142.

Van der Vleuten, C. P. M., Schuwirth, L. W. T., Scheele, F., Driessen, E. W., Hodges, B., Currie, R., et al. (2010). The assessment of professional competence: Building blocks for theory development. *Best Practice & Research. Clinical Obstetrics & Gynaecology, 24*, 703–719.

Weller, J. M., Jolly, B., Misur, M. P., Merry, A. F., Jones, A., Crossley, J. G. M., et al. (2009). Mini-clinical evaluation exercise in anesthesia training. *British Journal of Anaesthesia, 102*, 633–641.