Reliability of Mini-CEX Assessment of Medical Students in General Practice Clinical Attachments

Kyle Eggleton, MBChB, MMedSci, MPH; Felicity Goodyear-Smith, MBChB, MD; Lois Paton, MBChB(Hons); Karen Falloon, MBChB, PhD; Chris Wong, MBChB; Liza Lack, BM, BS, BMedSci, MRCGP; John Kennelly, MBChB; Tana Fishman, BS, MS, DO; Simon A. Moyes, MSc

BACKGROUND AND OBJECTIVES: Mini Clinical Evaluation eXercise (mini-CEX) involves observation of routine clinical encounters, initially developed to assess clinical competencies of postgraduate doctors. This study aimed to measure its inter-rater reliability in assessment of medical students in general practice settings.

METHODS: General practitioners (GPs) supervising medical students were invited to complete online teaching on how to conduct a mini-CEX. This included three randomly presented scripted films of clinical scenarios representing different levels of student performance. Consenting participants completed an online mini-CEX. Mean marks were calculated for each case, Intraclass Correlation Coefficients (ICC) for overall clinical and four individual competencies, one-way analysis of variance to compare scores, and internal consistency measured by Cronbach's alpha.

RESULTS: Results were collated for the first 100 completing GPs, majority aged 40–59 years (71%), male (59%), New Zealand European (58%). Forty-four percent were in rural practice, with 21 mean years in practice. Mean mini-CEX grades increased as standardized performance increased, indicating that GPs reliably agreed about ranking of student performance from poor to very good. The intraclass correlation coefficient (ICC) for overall clinical competency was 0.78 (95% confidence interval 0.48–0.99), indicating good reliability regarding their agreement with each other. A Cronbach's alpha calculated with the overall scores was 0.85, indicating good internal consistency.

CONCLUSIONS: Mini-CEXs in undergraduate general practice attachments provide a reliable measure of assessing performance. However, they may be less useful in identifying exceptional performance or weaknesses in key competencies. In addition, caution must be applied in relying upon mini-CEXs to supply a summative assessment.

(Fam Med 2016;48(8):624-30.)

The Mini Clinical Evaluation eXercise (mini-CEX) was developed by the American Board of Internal Medicine in response to concerns over the reliability and validity of using long cases to assess the clinical competence of postgraduate doctors (residents).[1] The encounters are designed to be short (15 to 20 minutes) and occur as a routine aspect of training. The assessor observes the trainee take a history, conduct a physical examination and offer a diagnosis and treatment plan, followed by feedback to the trainee at the end of the consultation.[2] Each resident should be evaluated multiple times by assessors in different hospital settings, which may be formative or summative.

Despite being commonly used, there are questions about the reliability and validity of this short assessment tool.[3] An early study of 88 residents assessed by mini-CEX, by 64 physicians over 355 encounters, found small variations in ratings (on the 9-point scale, ratings of overall clinical competence ranged from 5.5 to 8.0, and 45 of examiners (70%) had means between 6 and 7).[4] The tool was originally developed for

From the Department of General Practice and Primary Health Care, Faculty of Medical and Health Science, The University of Auckland, Auckland, New Zealand.

postgraduate medical settings, but it is now also used with undergraduate training, not only in medicine[5-7] but also in other disciplines, including dentistry[8] and nurse practitioners.[9] Construct and criterion validity of the mini-CEX were reviewed in a meta-analysis of 11 studies where mini-CEX was used to assess either medical residents' or students' (three studies[10-12]) clinical skills compared with other training measures.[3] This found a predictive validity coefficient ranging from 0.26 to 0.86.

A study of Year 1 postgraduate trainees found that eight mini-CEX per trainee were required to achieve a generalizability coefficient of 0.8 and that six assessments produced a coefficient of 0.75.[13] More encounters may be required when used for undergraduates. A study of its use on three occasions for each of five hospital attachments in medical students found the maximum reliability of 0.73 by aggregating the 15 scores.[6] A study of physicians rating performances of standardized residents on nine scripted clinical videotapes depicting three levels of performance (high satisfactory/superior, marginal/satisfactory, unsatisfactory) found that they were able to reliably discriminate between superior, satisfactory, and poor students, but there was a wide range in ratings for the same student among participants.[14]

To our knowledge there are only two studies looking at the use of mini-CEX in general practice, with both in postgraduate settings. One focused on the quality of written narrative feedback and reflection in a modified mini-CEX in GP trainees,[15] and the other assessed the reliability of its use as an assessment tool of practicing GPs.[16] In the latter, six raters scored 188 videotaped clinical encounters of 14 GPs with a generalisability coefficient of 0.92 for 10 encounters. However there is no research on the use of mini-CEX in undergraduate GP training.

In 2014 the University of Auckland introduced mini-CEX assessments for medical students (Year 5 of a 6-year program) in five hospital-based clinical attachments (surgery, medicine, obstetrics and gynaecology, pediatrics, psychiatry) plus general practice. In 2015 this was extended to Year 6 students. These assessments are standardized using a 4-point marking scale (excellent, satisfactory, some reservations, major deficiencies) for four competencies (history-taking/interviewing, physical examination skills, clinical judgement/reasoning, humanistic qualities/professionalism) plus an overall clinical competence grade of distinction: pass, borderline pass, or fail. In general practice, students have one or more formative and one final summative mini-CEX assessments conducted by the GP teacher during an attachment.

An online training module to train GPs on how to conduct these assessments was developed by a team of GP academics in the Department of General Practice and Primary Health Care. Instruction included the background of the mini-CEX, advice on how to score, including grade and competency descriptors and the marking rubric. Three case scenarios were scripted, each representing a different level of a Year 5 student performance—borderline pass, pass, or distinction (standardized cases 1, 2, and 3, respectively). Because the training module needed to fit into the standardized Royal New Zealand College of General Practitioners 1-hour duration for maintenance of professional standards, no scenario for fail was included. Films of each scenario were produced with roles played by a medical student, an academic GP, and three actors experienced in playing simulated patients and hosted on the GP teacher training website. For the physical examination component, snapshots were presented of the student conducting this, rather than filming the entire sequence. Each film was randomly presented to the GP teachers undergoing the training, who completed the online mini-CEX form before viewing the next film. Separate films of the academic GP giving the student constructive feedback on their performance once the patient has departed then were presented. At the end of the session, consensus marks allocated to each case by the team who had scripted the scenarios were presented online.

There is a paucity of information about the reliability of mini-CEX in undergraduate students and particularly with its use outside hospital settings. The aim of this study was to use our training module to assess its inter-rater reliability in the general practice assessment of medical students. Our hypothesis was that mini-CEX would have good reliability in an undergraduate general practice setting. Specifically, we wanted to know whether raters reliably agreed about the ranking of student performance from poor to good, how reliably they agreed with each other, and how reliably they agreed with our "official" rating of a performance.

Methods

Participants

All GPs supervising University of Auckland Year 5 or Year 6 medical students during their community placements (approximate number 200) were asked to undertake the online training on how to conduct a mini-CEX assessment. When invited by email to register with the website, and complete the teaching module, they were also provided with a Participant Information Sheet and asked if they would enroll in the study (meaning that their assessment data would be captured for analysis). Declining GPs could still undertake the training. Academic GPs in the Department of General Practice and Primary Health Care who also supervised students in their practices were excluded. Participants entered their demographic information into the study website, then viewed the films in random order once they had undertaken training. All GPs scored the case electronically using the mini-CEX rubric after watching each film, before progressing to the next one.

Ethical Approval

Approval was granted by the University of Auckland Human Participants Ethics Committee, Ref 01116, March 19, 2014.

Sample Size

We had previously conducted a pilot study at a face-to-face training session of 24 GP teachers. All GPs watched a role-played mock case with a student interviewing a patient supervised by a GP, then marked the student using the mini-CEX. The intraclass correlation coefficient (ICC) between assessors was 0.38 (0.18, 0.79). Randomly resampling these data, we calculated that a sample of 50 GPs could reduce the 95% confidence interval of the ICC down to about 0.59 from the observed 0.61. Because the pilot group may include more GP teachers interested in teaching or who have had previous training in medical education than the larger group of all teaching GPs, they may be more homogenous with respect to their student assessment skills. The sample size was increased by 100% to 100, to adjust for the potential increase in variability among GP teachers. While the research component closed at 100, the training component continues to be available.

Analyses

While in a mini-CEX, assessors grade the student with respect to history-taking, physical examination, clinical judgment, and humanistic qualities, they mark the overall clinical competency of the student as distinction, pass, borderline pass, or fail based on these plus a global assessment of the student's performance. Comparing each individual competency may be problematic, because of the "halo" effect, in which competencies influence the scoring of each other.[17] The overall competency grade was scored as distinction=4, pass=3, borderline pass=2, fail=1. Mean marks, medians, and range were calculated for each scenario. The ICC was calculated for the overall competency. The Intraclass Correlation Coefficient (ICC) is a measure of similarity between measurements. If there is perfect consistency across all GP teachers, ie, each GP teacher scored the mini-CEX in exactly the same way, the ICC would be 1. If there seems to be no relationship at all between GP teachers then the ICC would be 0. Generally a ICC > 0.7 represents acceptable consistency between reviewers. The ICC calculations in SAS used a SAS macro based on the method introduced by Shrout and Fleiss.[18] This was double checked with another SAS macro written by Hays, Wang, and Sonksen.[19] Analysis of variance (ANOVA) was used to calculate Kendall's Coefficient of Concordance to compare ordinal scores and Kappa statistics for nominal responses. Internal consistency was measured by Cronbach's alpha.[20] Calculations were conducted using SPSS 21 and SAS 9.3.

Results

Of the GP teachers who were approached and undertook the website-based training for marking a mini-CEX, 122 consented to participating and 100 completed the study. Following entry of 100 participants into the study, the teaching module continued but data collection ceased. Their demographic characteristics are presented in Table 1. The majority (71%) were aged between 40 and 59 years, male (59%), New Zealand European (58%), 44% were in rural practice and with 21 mean years in practice.

Mean Grades and Inter-Rater Reliability

Table 2 shows the mean, median, and range of scores for overall performance for each case. The mean of the mini-CEX increased as the standardized performance increased, indicating that the GPs reliably agreed about the ranking of student performance from poor to very good.

The intraclass correlation coefficient (ICC) for overall clinical competency was 0.78 (95% confidence interval 0.48–0.99), indicating good reliability regarding their agreement with each other. Kendall's Coefficient of Concordance for ordinal response was 0.8 (see Table 3). A Cronbach's alpha calculated with the overall scores was 0.85, indicating good internal consistency. The proportion of GPs who gave specific scores for each case can be seen in Figure 1, and comparison to the standardised scores can be made. There was one outlier who graded all three cases as fail.

Discussion

The ICC of 0.78 indicates that the GP teachers show relative consistency between them in scoring the mini-CEX. However, there was a wide range of marks, with a range of 1–3 (fail) or 1–4 (borderline pass, pass, distinction), indicating the presence of outlier assessors.

On average, participants judged the student performance lower than the overall grade considered appropriate for each case by the academic team. However, use of filmed scenarios may not fully represent real-life teaching. Students in general practice settings have a one-on-one relationship with their GP teacher for several weeks, and in these circumstances they may judge them less harshly than an anonymous student viewed on film. Another possible reason for a difference between academic GPs and GP teachers, in regards to the overall grade, is that the academic GPs based their grade on the scripts that were devised for each scenario rather than the video. Social judgements based on appearance may alter a rater's perception of a student and may also account for the small degree of variability seen in this study. For example a number of GP teachers commented on the untidy hair of the student in the videos. Evidence suggests that social judgement can account for between 9%–57% of the variance seen in mini-CEXs.[21]

Variability between raters may also be due to individual raters unwillingness to use labels such as "unsatisfactory" for trainees because

Table 1: Demographics of Participants

Demographic	Number
Age	
20–29	4
30–39	8
40–49	23
50–59	48
60–69	17
Gender	
Males	59
Females	41
Ethnicity	
New Zealand European	58
Other European	17
Indian	13
Chinese	5
Other Asian	4
Samoan	2
Decline	1
Mean years teaching	7.6
Mean years as GP	21.2
Rural GP	44

Table 2: Mean, Median, and Range of Scores for Each Case

Case 1 Borderline Pass			Case 2 Pass			Case 3 Distinction		
Mean	Median	Range	Mean	Median	Range	Mean	Median	Range
1.4	1	1–3	3.1	3	1–4	3.4	3	1–4

Where fail=1, borderline pass=2, pass=3, distinction=4

they perceive them as pejorative. Use of construct-aligned scales (such as "performed at the level expected," "performed at a higher level than expected," "performed at a lower level than expected") has been shown to significantly reduce assessor disagreement, with the number of mini-CEX assessors required to achieve a generalizability coefficient >0.70 falling from six to three.[22]

A potential weakness of this study is that it does not report on individual competencies. A decision was made to not analyse individual competencies because of the "halo" effect[17] as well as our primary interest being the reliability of the rating of the overall performance. A review of the literature on mini-CEX found that individual competencies are highly inter-correlated and that this might make it difficult for raters to discriminate between individual strengths and weaknesses of students.[12,21] Other literature has also suggested that individual competencies provide only a minimal contribution to overall score in an undergraduate setting.[6] Assessors may be able to distinguish between overall excellent, good, and fair performances by students, but the exact rating they give each competency to reach that grade may vary considerably between markers.

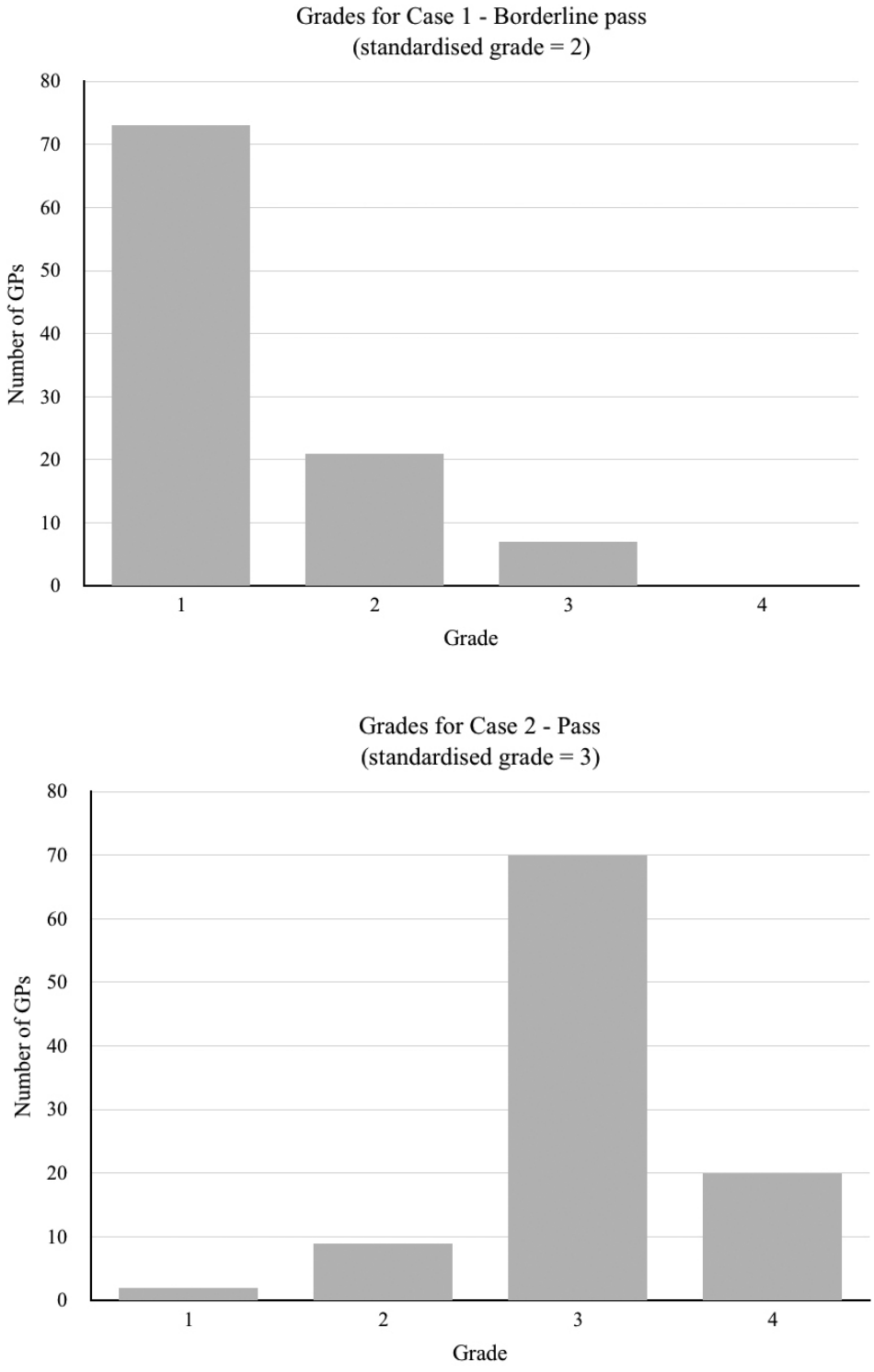
Strengths of this study included achieving the required sample size (100) and the random presentation of the three scenarios to reduce within-rater bias.

Mini-CEX became part of the assessment for Year 5 medical students at the University of Auckland in 2014 and has been introduced for Year 6 students in 2015 (following data collection for this study). Training and familiarity with mini-CEX might improve the accuracy of marking. However, evidence is mixed on the effect of training. One study found no significant difference in mini-CEX scores between trained and untrained raters,[23] whereas a randomised controlled trial using scripted videos found statistically significant differences in medical interviewing and physical examination ratings between intervention and control groups.[24] Training may, however, increase the level of

Table 3: Estimate of Variability

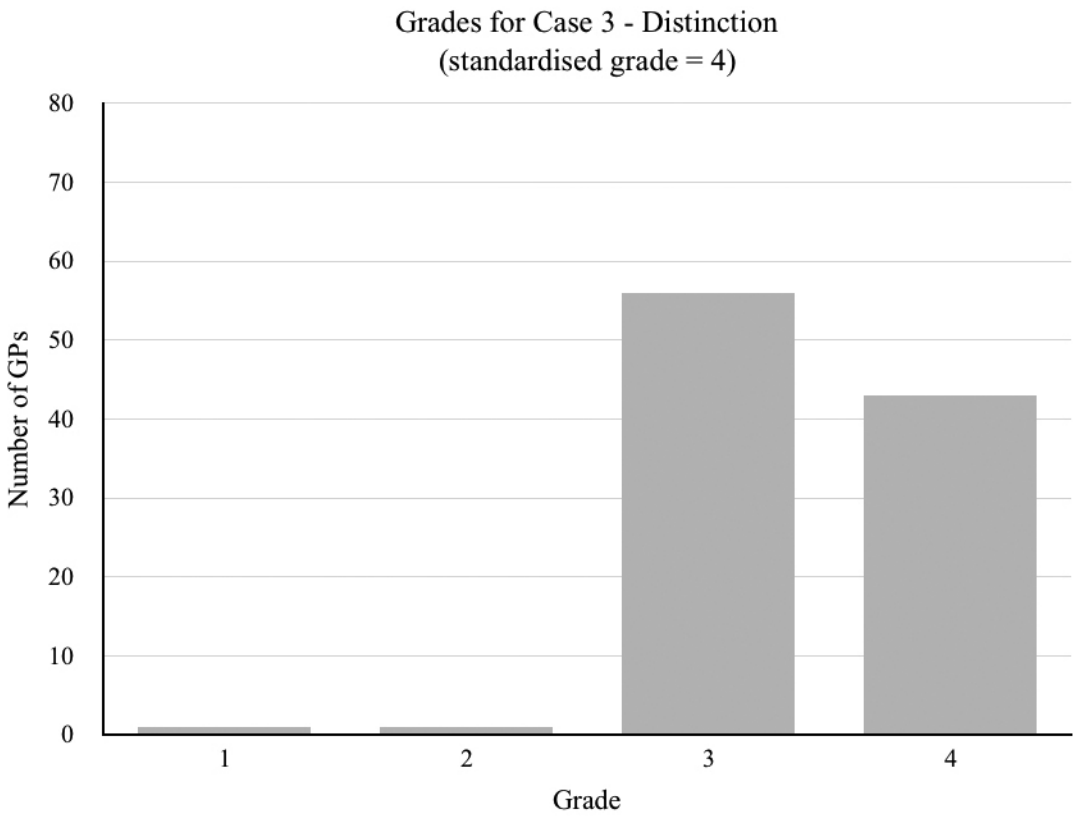
	ICC ICC (2,1)	Reliability ICC (2,100)	Kappa Statistics for Nominal Response	Kendall's Coefficient of Concordance for Ordinal Response
Overall Grade	0.775 [0.481, 0.993]	0.997 [0.989, 1.000]	0.313	0.806

Figure 1: Proportion of GPs Assigning Grades to Each Case



(continued on next page)

Figure 1: continued



confidence and comfort for teachers in scoring a mini-CEX.[25]

Judgement of what is considered a good performance for a trainee also needs to take into account their level of training (Year 6 students should be performing at a higher level than Year 5), and there may also be an effect of time of the year. A study of postgraduate Year 1 trainees found the critical judgement and organization scores increased significantly over time, suggesting trainee improvement with time.[13]

Increasingly, Mini-CEXs are becoming established as a way of assessing undergraduate and postgraduate students. This study of undergraduate GP teachers suggests that the mini-CEX can provide a reliable measurement of overall performance. However, caution must be applied in relying too heavily on the results of one mini-CEX to determine summative performance. This is predominately due to the presence of outlier assessors and the small degree of variability between assessors.

However, mini-CEXs do not serve solely as summative assessments. GPs are encouraged to conduct several formative mini-CEXs prior to the final graded one. These mini-CEX, if combined with interventions to improve the educational feedback of GP teachers, can provide an opportunity for students to reflect on their practice immediately after a clinical encounter, which can be a powerful learning tool.[15]

ACKNOWLEDGMENTS: This research was funded by the E.W. Sharman Staff Award for Curriculum Development, University of Auckland.

CORRESPONDING AUTHOR: Address correspondence to Dr Eggleton, Department of General Practice and Primary Health Care, Faculty of Medical and Health Science, The University of Auckland, Auckland, New Zealand. +64 21 686 487. Fax: +64 9 373 7624. k.eggleton@auckland.ac.nz.

References

1. Norcini JJ, Blank LL, Arnold GK, et al. The mini-CEX (clinical evaluation exercise): a preliminary investigation. Ann Intern Med 1995;123(10):795-9.
2. Norcini JJ, Blank LL, Duffy FD, et al. The mini-CEX: a method for assessing clinical skills. Ann Intern Med 2003;138(6):476-81.
3. Al Ansari A, Ali SK, Donnon T. The construct and criterion validity of the mini-CEX: a meta-analysis of the published research. Acad Med 2013;88(3):413-20.
4. Norcini JJ, Blank LL, Arnold GK, et al. Examiner differences in the mini-CEX. Adv Health Sci Educ Theory Pract 1997;2(1):27-33.
5. Jabeen D. Use of simulated patients for assessment of communication skills in undergraduate medical education in obstetrics and gynaecology. J Coll Physicians Surg Pak 2013;23(1):16-9.
6. Hill F, Kendall K, Galbraith K, et al. Implementing the undergraduate mini-CEX: a tailored approach at Southampton University. Med Educ 2009;43(4):326-34.
7. Fernando N, Cleland J, McKenzie H, et al. Identifying the factors that determine feedback given to undergraduate medical students following formative mini-CEX assessments. Med Educ 2008;42(1):89-95.

8. Behere R. Introduction of Mini-CEX in undergraduate dental education in India. Educ Health 2014;27(3):262-8.
9. Milner KA, Watson SM, Stewart JG, et al. Use of Mini-CEX tool to assess clinical competence in family nurse practitioner students using undergraduate students as patients and doctoral students as evaluators. J Nurs Educ 2014;53(12):719-20.
10. Torre DM, Simpson DE, Elnicki DM, et al. Feasibility, reliability and user satisfaction with a PDA-based mini-CEX to evaluate the clinical skills of third-year medical students. Teach Learn Med 2007;19(3):271-7.
11. Kogan JR, Bellini LM, Shea JA. Feasibility, reliability, and validity of the mini-clinical evaluation exercise (mCEX) in a medicine core clerkship. Acad Med 2003;78(10 Suppl):S33-5.
12. Boulet JR, McKinley DW, Norcini JJ, et al. Assessing the comparability of standardized patient and physician evaluations of clinical skills. Adv Health Sci Educ Theory Pract 2002;7(2):85-97.
13. Jackson D, Wall D. An evaluation of the use of the mini-CEX in the foundation programme. Br J Hosp Med (Lond) 2010;71(10):584-8.
14. Holmboe ES, Huot S, Chung J, et al. Construct validity of the miniclinical evaluation exercise (miniCEX). Acad Med 2003;78(8):826-30.
15. Pelgrim EA, Kramer AW, Mokkink HG, et al. Quality of written narrative feedback and reflection in a modified mini-clinical evaluation exercise: an observational study. BMC Med Educ 2012;12:97.
16. Sidhu RS, Hatala R, Barron S, et al. Reliability and acceptance of the mini-clinical evaluation exercise as a performance assessment of practicing physicians. Acad Med 2009;84(10 Suppl):S113-5.
17. Schneider F, Gruman J, Coutts L. Applied social psychology. Thousand Oaks, CA: Sage Publications Inc, 2012.
18. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 1979;86(2):420-8.
19. Hays RD, Wang E, Sonksen M. General reliability and intraclass correlation program (GRIP). Proceedings of the 3rd Annual Western Users of SAS Conference, Long Beach, CA, 1995:220-3.
20. Cronbach L. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297-334.
21. Hawkins RE, Margolis MJ, Durning SJ, et al. Constructing a validity argument for the mini-Clinical Evaluation Exercise: a review of the research. Acad Med 2010;85(9):1453-61.
22. Crossley J, Johnson G, Booth J, et al. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. Med Educ 2011;45(6):560-9.
23. Cook DA, Dupras DM, Beckman TJ, et al. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. J Gen Intern Med 2009;24(1):74-9.
24. Arora VM, Berhie S, Horwitz LI, et al. Using standardized videos to validate a measure of handoff quality: the handoff mini-clinical examination exercise. J Hosp Med 2014;9(7):441-6.
25. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. Ann Intern Med 2004;140(11):874-81.