Multifaceted Assessment in a Family Medicine Clerkship:

A Pilot Study

Valerie Dory, MD, MMedEd, PhD; Bernard Charlin, MD, PhD; Dominique Vanpee, MD, PhD; Robert Gagnon, MPsyc

BACKGROUND AND OBJECTIVES: Programs of assessment should reflect the multifaceted nature of medical competence. We experimented with new testing methods, ie, script concordance testing (SCT) and clinical reasoning problems (CRPs), combined with the habitual OSCE for an end of family medicine clerkship. Our aims were to compare students' scores with experts' scores, to determine whether the new tests detected learning over a 3-month period, and to examine whether the tests were redundant.

METHODS: We conducted a longitudinal study on one cohort of family medicine clerks. Two formative testing sessions using both SCT and CRPs were held 3 months apart. Students' scores were compared to those of the panel of experts used to score the tests. We examined the difference in students' scores between the two testing sessions. Finally, we computed correlation coefficients between these scores and the summative OSCE.

RESULTS: Panelists' scores were significantly higher than students' scores. SCT scores did not change significantly over 3 months whereas CRP scores improved (Wilcoxon z -3.058, effect size 0.461, *P*=.002). Correlations between the OSCE and the written tests were low or non-significant. There were low correlations between the first CRP and both SCTs (Spearman's rho 0.357 and 0.358) but not between the second CRP and any SCT.

CONCLUSIONS: Written tests of clinical reasoning could provide relevant additional information to the evaluation of students' competence over the course of a family medicine clerkship. Further research is needed to determine the potential educational consequences of such programs of assessment.

(Fam Med 2014;46(10):755-60.)

Clinical competence is multifaceted and requires the integration of knowledge, skills, and attitudes.[1] Its assessment also requires an integrated and multifaceted approach.[2] Assessment of clinical competence should use a variety of instruments to provide a 360° picture of each learner's development.[2] However, if one accepts that competence is not just about possessing the required knowledge, skills, and attitudes but is more about integrating them in practice, assessment should not solely focus on assessing the components of competence but should include occasions where students can demonstrate their ability to use them together effectively.[1]

Like in other family medicine clerkships, students at our institution are assessed using a summative Objective Structured Clinical Examination (OSCE).[3] The OSCE provides tasks requiring students to demonstrate the integration of knowledge, skills, and attitudes. However, by targeting complex performance, it is somewhat limited in the detail it can provide about specific components of the performance, which could be important for feedback purposes. This is particularly true regarding the process of clinical reasoning where the process is largely invisible and can only be inferred from certain behaviors exhibited during the OSCE (eg, the questions an examinee asks a simulated patient can reflect their underlying diagnostic hypotheses).[4] Further, the case-specificity of clinical reasoning requires wide sampling of material for reliable judgments to be made.[5] We therefore conducted a pilot study to examine the utility of adding other types of assessments, specifically targeted at clinical reasoning, during our family medicine clerkship.

Because our concern was specifically about students' ability to

From the Institute of Health and Society (IRSS), Université catholique de Louvain, Brussels, Belgium (Drs Dory and Vanpee); Fonds de la Recherche Scientifique–FNRS, Brussels, Belgium (Dr Dory); Undergraduate Medical Education and Centre for Medical Education, Faculty of Medicine, McGill University (Dr Dory); and Faculty of Medicine, Université de Montréal (Drs Charlin and Gagnon).

reason through family medicine cases, we chose two written tests of clinical reasoning, ie, the script concordance test (SCT) and clinical reasoning problems (CRPs). The script concordance test is designed to assess clinical data interpretation in situations of uncertainty.[6] It is highly efficient, providing reliable scores within very reasonable testing times (around 60–90 minutes).[6] It is also particularly valuable for testing reasoning in ill-defined problems, unlike traditional single-best answer multiple-choice questions.[7] Learner responses are compared with those of a panel of experts, and partial credit is awarded based on the number of experts selecting a specific response, a practice referred to as aggregate scoring.[6] The SCT presents examinees with a brief, ill-defined, clinical scenario. Each scenario is followed by three to four questions, which include a suggested hypothesis and a new piece of information. The question pertains to the impact of the new piece of information on examinees' evaluation of the suggested hypothesis (see Table 1). Although some argue that it is one of the SCT's strengths, the limited focus of the SCT on clinical data interpretation led us to consider adding a different test of clinical reasoning that probes other facets of clinical reasoning.

CRPs present learners with longer clinical scenarios with a certain amount of uncertainty, ie, there is no single solution to the case.[8] Examinees are required to provide two diagnoses in free-text format and to select and weigh cues from the clinical scenario in terms of their contribution to the diagnostic hypotheses proposed (Appendix available from the corresponding author by request). Like the SCT, CRP scores are derived from the responses of a panel of experts. Unlike the SCT, CRPs are not limited to measuring clinical data interpretation but also provide an evaluation of learners' ability to generate hypotheses.[9] Indeed, the CRP marking scheme provides an overall score and two subscores, one for the quality of the diagnostic hypotheses generated (diagnostic subscore) and one for the identification and weighing of key clinical features supporting/disconfirming the proposed hypotheses (features subscore). Tests comprising 10 clinical scenarios have been shown to provide reasonably reliable scores (alphas of approximately 0.70) within estimated testing times of 90–120 minutes and to detect differences between student levels within a medical curriculum.[8]

The aims of our study were to pilot the addition of SCT and CRP to the clerkship's assessment. In view of the pilot nature of the study, the additional tests were not used as part of the summative assessment and served a purely formative purpose (ie, scores on the SCT and CRP were used for feedback, they were not included in students' grades and did not influence pass-fail decisions). Our first goal was to describe the scores obtained by students on the additional tests and to compare them to those of a panel of expert family physicians. We hypothesized that the tests would discriminate between experts and students, ie, that experts would obtain significantly higher scores than students. We also strove to explore whether the new tests would be sensitive enough to detect the learning occurring during the clerkship itself by looking at students' progress over a 3-month period. Finally, we sought to examine whether the various tests were redundant by studying the intercorrelations of test scores. We hypothesized that the correlations between the SCT and CRP would be moderate (with higher correlations between the SCT and the key features subscore of the CRP as compared with the diagnostic subscore)

Table 1: Sample Script Concordance Test (SCT) Item Containing Four Questions

Case 1: You are on duty one night when you are called for a home visit for Nora, aged 22. She has been suffering from severe pain in the right iliac fossa for several hours now. She has vomited once.							
	If you were thinking . . .	And you discover that . . .	The hypothesis becomes . . .				
			-2	-1	0	+1	+2
1.1	Ectopic pregnancy	There is guarding in the right iliac fossa					
1.2	Ovarian torsion	She takes her hormonal contraception regularly					
1.3	Appendicitis	She's had one loose stool					
1.4	Kidney stones (ureteral colic)	Her axillary temperature is 38.5° C					

-2 = much less likely (or ruled out)
-1 = less likely
0 = no change in likelihood
+1 = more likely
+2 = much more likely (or totally confirmed)

and that correlations between these tests and the OSCE would be low.

Methods

Context

In the current 7-year curriculum, those electing to specialize in family medicine spend their final semester of medical school in family medicine clerkships 4 days a week with 1 day a week on campus for small-group teaching. The semester is divided into six 1-month periods, but students usually stay in the same family practice for two or three periods. Students can elect to spend one period in an alternative community setting. A minority of students opt to do so. The summative assessment consists in an OSCE and several assignments (ie, for the evidence-based medicine course, for a course on practice management and a reflective assignment; we did not examine data from assignments in this study). The year the study was conducted, the OSCE had 14 stations. Ten stations were 7.5 minutes long, and four were 15 minutes long. Their themes were often linked to the content of small-group teaching sessions.

Participants and Procedure

We presented the aims and methods of the study to students during one of their on-campus days at the beginning of the semester in January. All students enrolled in the program were eligible. We asked volunteers to sign a consent form.

We asked volunteers to take part in two formative distance assessment sessions, 3 months apart, at the beginning of February and at the beginning of May. Students could provide written comments at the end of the first test form. Some students volunteered comments by email.

Students received their scores and information on the scores of the group of participants by email. A feedback session was organized at the end of May to discuss test answers.

We also asked participants to grant us access to data from their summative OSCE assessment.

Participants were given a stipend of 10 euros (approximately 14 US $), and their names were entered into a raffle for three gift vouchers worth 150 euros each (approximately 200 US $).

Tests

Two tests were constructed, each comprising 20 SCT cases and 10 CRP cases, for an estimated testing time of 3–4 hours per test. The same blueprint was used for both tests. It included 30 clinical presentations, of which seven concerned children and teenagers, eight older patients, and two women's health issues. Presentations were selected to cover a broad range of systems and to be representative of family practice.

SCT cases were written by the first author. The 10 CRP cases used in the first testing session were translations of cases developed by Groves and colleagues (personal communication). The first author wrote the 10 CRPs used in the second testing session. She interviewed family medicine and geriatrics faculty members to explore common presentations and their key clinical features.[10] The resulting questions were presented to family medicine faculty for their opinions on relevance, authenticity, and ease of comprehension. The tests were prepared in PDF format and were sent to participants by email. Participants had 3 days to complete the tests. They were instructed not to use any external help such as advice from peers or supervisors, the internet, or textbooks.

We recruited 16 experts, all family physicians, of which eight were faculty members at our institution, one was a member of another institution's faculty, and four were clinical supervisors. They completed both tests in their own time. One expert only provided responses for the first test.

SCTs were scored using the Excel calculator available on its developers' website (http://www.cpass.umontreal.ca/sct.html). CRP free-text responses were converted into standard categories decided upon by the first author. Scoring keys based on experts' responses were computed manually for CRPs. Keys and standardized responses were used to compute scores using software developed at the University of Queensland where CRPs were initially developed.

The first SCT had an internal consistency of alpha 0.79, the second SCT yielded a lower alpha of 0.66. The first CRP had low internal consistency (alpha 0.49), whereas the second had an acceptable alpha of 0.76.

Analyses

We used non-parametric statistics because not all data were normally distributed. We compared student and expert scores using the Mann-Whitney test and computed an effect size r.[11]

We calculated participants' progress by first standardizing the scores according to the panel's mean scores (yielding so called centered scores).[12] We compared scores on both testing occasions using Wilcoxon's signed-rank test and calculated effect size r as described above. We examined the relationship between the change of centered scores and clerkship experience data using Spearman correlations. We computed Spearman correlations between scores.

All analyses were conducted using IBM SPSS Statistics version 20.

Ethical Considerations

The study was granted ethical approval from the faculty's ethics committee. Participation was voluntary, and the names of participants were kept confidential and were specifically not disclosed to family medicine faculty or clinical supervisors.

Results

Although 55 of 59 students gave consent (consent rate 93%), participation

waned over the course of the 3 months of the study. Forty-four students took part in both testing sessions (participation rate 75%). Only their data were analyzed.

Experts scored significantly better than students on all tests (Figure 1). The effect sizes were moderate for the SCT (0.48 and 0.45 for tests 1 and 2 respectively) and large for the CRP (0.73 and 0.67), particularly for the features subscore of the CRP (0.76 and 0.73 for the features subscores versus 0.36 and 0.27 for the diagnoses subscores).

Students had significantly higher centered CRP scores in the second testing session, but their SCT scores did not change significantly (Figure 2).

Both SCT tests were slightly correlated with the first CRP test (Spearman's rho 0.357 and 0.358, *P* values both 0.017) but not the second (Spearman's rho 0.126 and 0.059, *P* values 0.413 and 0.705). We did not find the expected pattern of CRP features subscores correlating more strongly with SCT than CRP diagnoses subscores. There were low correlations between the OSCE and SCT although the result was not statistically significant for the second SCT. The CRPs were not correlated with the OSCE.

We received comments from 20 students. Negative comments included finding the tests too long (n=4; one student specified having spent around 4 hours for the first test), finding the tests difficult (n=3), finding them not representative of family medicine (n=1). One student commented on feeling frustrated by not having an opportunity to justify responses. One student found it difficult to distinguish between +/- 1 and 2 on the Likert scale for the SCT. Two students found it difficult to provide more than one hypothesis on the CRP, whereas one student found it difficult to select only two hypotheses from his/her differential. One student was unsure about whether the hypotheses generated for the CRP should only include those that were deemed likely or whether one should include diagnoses that one was seeking to exclude. One student found it difficult to weigh data in the CRP. Positive comments included finding the tests interesting (n=6). One student stated that completing the tests had prompted her/him to study more. One student was positive about having enough time to complete the written tests in comparison to the OSCE where they had felt overwhelmed by the time pressure.

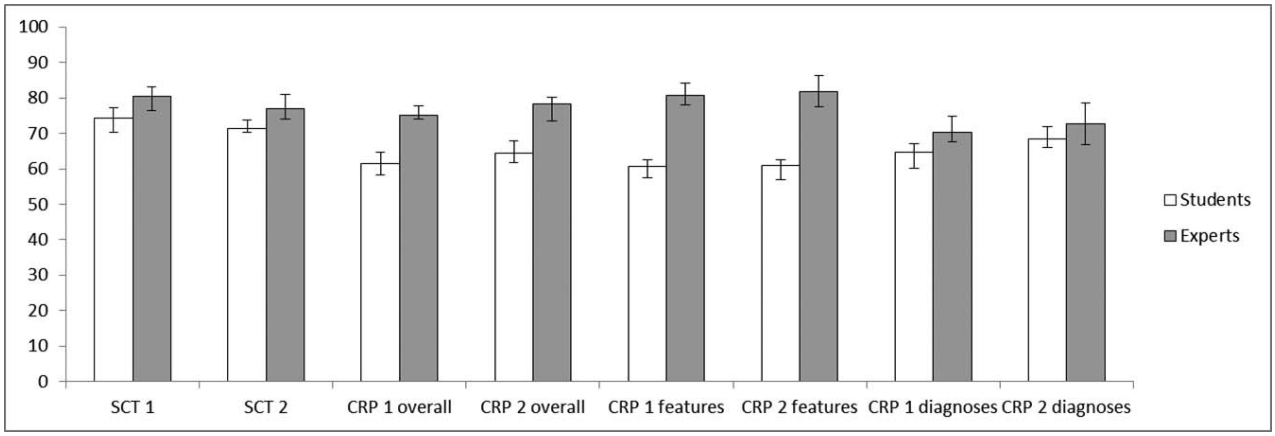
Discussion

Main Findings

This study piloted the addition of two written tests of clinical reasoning to an existing OSCE as part of an assessment program for a family medicine clerkship.

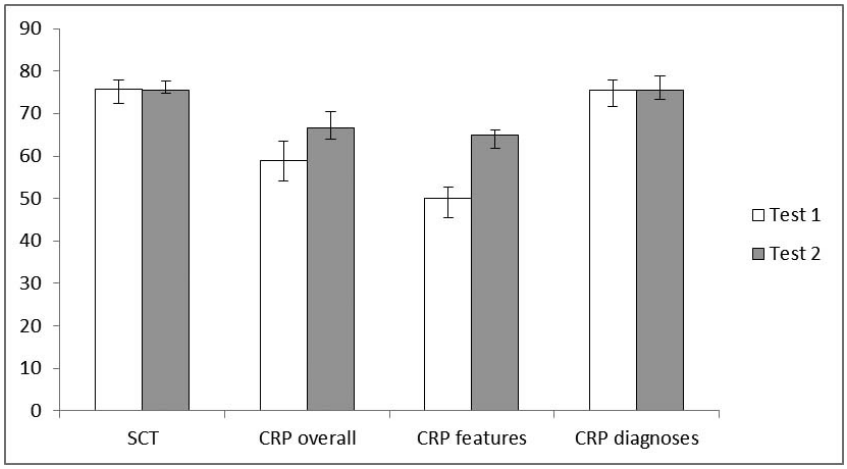
Tests examined individually had strengths and weaknesses. Both tests were able to clearly distinguish between clerks and family physicians. The SCT provided more reliable scores than did the CRP. This is not surprising in view of the fact that there were twice as many SCT cases as CRP cases, which is a determining factor in a test's reliability. Because of the longer format of CRPs, CRPs are less efficient in this respect, and longer CRP tests would become much less feasible. The CRP on the other hand was more sensitive to changes occurring over a 3-month period within the clerkship. To our knowledge no other studies have examined the CRPs' sensitivity to change over such a short period of time. As for the SCT, a few studies have previously documented improvements in SCT scores over the course of clerkship rotations. Two small studies using the same test before and after a rotation, thus compounding the potential impact of the rotation and the actual testing itself,[13] have indicated significant improvements.[14,15] One group-control

Figure 1: Students' and Experts' Median Test Scores and Subscores With 95% Confidence Intervals



SCT— script concordance test
CRP— clinical reasoning problems

Figure 2: Students' Performance on the Two Testing Occasions*



* Bars represent median centered scores and subscores with their 95% confidence intervals. Centered scores are standardized scores based on the reference panel score distribution.

study also found significantly higher scores on a nephrology SCT in students who had taken part in a nephrology rotation.[16]

While more reliable, the SCT should not supplant the CRP since both appear to be measuring somewhat different components of clinical reasoning, as evidenced by the low correlations found between scores on the two in this study. Although both were designed to evaluate clinical reasoning, the SCT focuses exclusively on the interpretation, ie, weighing, of clinical data in light of suggested hypotheses, whereas the CRP requires examinees to generate hypotheses, identify key features of a case, and then weigh these features. From our data, it appears that these distinctions lead to significant differences in examinees' results, which clearly supports the use of both within an assessment program.

As expected, the correlations between the written tests and the OSCE were low at best. Another study compared the SCT and OSCE scores 2 years later and found similar results.[17] While clinical reasoning is involved in many OSCE tasks, the OSCE also measures other aspects of clinical competence such as communication and technical skills. As such it provides a broader overview of students' medical competence. We would nevertheless argue that, because it measures clinical reasoning indirectly, ie, through outputs such as the questions examinees ask simulated patients and because of the limited number of cases proposed in an OSCE, it is limited in the depth of exploration of clinical reasoning, which is a key component of medical competence. Introducing a written test including SCT and CRP therefore provides additional information regarding students' clinical reasoning ability, which is important for feedback purposes. By covering a broader scope of clinical presentations, this would also compensate the limited sampling of an OSCE.

Limitations

Our participation rate was high, but the study was conducted in a single institution, which limits the generalizability of our findings. The relatively small sample of participants may have led to a lack of power of the study. Further, the relatively low reliability of the CRP, particularly in the first testing session, is likely to have attenuated the correlations between CRP scores and scores on other tests. Finally, students completed the SCT and CRP on a voluntary basis and at home, with no supervision. This could have reduced their cognitive engagement with the task and may have led some students to use external resources (eg, textbooks, internet, friends) to respond. This could explain some of the divergence observed between these test scores and those on the OSCE, which was completed in a stringent, summative, context. The high participation rate and informal communications indicate that students were generally highly motivated in spite (or perhaps because) of the formative nature of the task. We have no data regarding the use of external resources.

Educational Implications

Despite the limitations of a single-site single-cohort study, our study contributes to the growing evidence pertaining to the SCT[18] and CRP.[8,9,19] The SCT has been widely used since its development in the late 1990s. Our findings confirm its internal consistency and discriminating ability but failed to show sensitivity to change over a 3-month period of clinical clerkship. Other studies have found positive evidence in this regard.[14-16] It may be that family practice is such a broad domain that significant change requires more time although students' did improve their CRP scores.

CRPs have not yet been widely used, probably because of the complexities of scoring free-text responses. Since our study, the CRP's developers have designed an online platform allowing automated scoring of CRPs without the manual preparation required in this study. This may lead to wider adoption of the CRP. CRPs probe several aspects of clinical reasoning, including the ability to generate (rather than select) diagnostic hypotheses. Our findings further suggest that the CRP provides a distinct perspective on students' competence and that it is sensitive to change following a relatively short span of clinical teaching.

We did not examine the educational consequences of adding these tests. Nevertheless, the high participation rate and a number of comments from students indicate that students found the tests useful. Further research is needed to examine how implementing changes in assessment programs might steer students' learning in preparation for the

tests[20] and how feedback from tests could be used to further learning.[21]

Suggestions for Further Research

Current understandings of assessment of clinical competence have led to a shift away from seeking the silver bullet instrument to measure individual components of competence.[2] This provides opportunities as well as challenges. On the one hand, by using multiple methods, assessment programs can in some respects compensate the inherent weaknesses of individual assessment methods. On the other, this raises the question of how to ensure and evaluate the quality of an assessment program as a whole. Our study combined assessment methods and sought to analyze the utility of combining them in terms of the correlations between various scores. While this approach is useful, it cannot provide a comprehensive and holistic assessment of the quality of an assessment program. Efforts in developing methods to do this are required before further research is conducted in this area.[21]

ACKNOWLEDGMENTS: At the time of the study, Valérie Dory was a post-doctoral researcher financed by the Fonds de la Recherche Scientifique–FNRS.

The authors thank the students and experts for their time and effort in completing the tests. Thanks to Michele Groves, University of Queensland, for providing assistance in the development and scoring of CRPs and for allowing us to use some from her item bank. Thanks to Stefan Maetschke, University of Queensland, for his assistance in the computerized scoring of CRPs.

CORRESPONDING AUTHOR: Address correspondence to Dr Dory, Lady Meredith House, 1110 Pine Avenue West, Montreal, Quebec H3A 1A3, Canada. 514-398-4400 ext 00866. Fax: 514-398-7246. valerie.dory@mcgill.ca.

References

1. Fernandez N, Dory V, Ste-Marie LG, Chaput M, Charlin B, Boucher A. Varying conceptions of competence: an analysis of how health sciences educators define competence. Med Educ 2012;46(4):357-65.
2. van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. Med Educ 2005;39(3):309-17.
3. Prislin MD, Fitzpatrick CF, Lie D, Giglio M, Radecki S, Lewis E. Use of an objective structured clinical examination in evaluating student performance. Fam Med 1998;30(5):338-44.
4. Kreiter CD, Bergus G. The validity of performance-based measures of clinical reasoning and alternative approaches. Med Educ 2009;43(4):320-5.
5. Norman G. Research in clinical reasoning: past history and current trends. Med Educ 2005;39(4):418-27.
6. Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: from theory to practice: AMEE Guide No. 75. Med Teach 2013;35(3):184-93.
7. Elstein AS. Beyond multiple-choice questions and essays: the need for a new way to assess clinical competence. Acad Med 1993;68(4): 244-9.
8. Groves M, Scott I, Alexander H. Assessing clinical reasoning: a method to monitor its development in a PBL curriculum. Med Teach 2002;24(5):507-15.
9. Groves M, O'Rourke P, Alexander H. Clinical reasoning: the relative contribution of identification, interpretation and hypothesis errors to misdiagnosis. Med Teach 2003;25(6):621-5.
10. Fournier JP, Demeester A, Charlin B. Script concordance tests: guidelines for construction. BMC Med Inform Decis Mak 2008;8:18.
11. Rosenthal R, DiMatteo MR. Meta-analysis: recent developments in quantitative methods for literature reviews. Annu Rev Psychol 2001;52(1):59-82.
12. Charlin B, Gagnon R, Lubarsky S, et al. Assessment in the context of uncertainty using the script concordance test: more meaning for scores. Teach Learn Med 2010;22(3):180-6.
13. Larsen DP, Butler AC, Roediger HL III. Test-enhanced learning in medical education. Med Educ 2008;42(10):959-66.
14. Joly L, Braun M, Fournir JP, Benetos A. Test de concordance de script et apprentissage du raisonnement clinique en gériatrie : intérêt dans la formation en stage des étudiants hospitaliers. Pédagogie Médicale 2009;10(Suppl 1):S39.
15. Gibot S, Bollaert PE. Le test de concordance de script comme outil d'évaluation formative en réanimation médicale. Pédagogie Médicale 2008;9(1):7-18.
16. Sqalli Houssaini TS, Bono W, Tachfouti N, Maillard D. Pertinence d'un test de concordance de script dans l'évaluation des compétences en néphrologie des étudiants du deuxième cycle de la faculté de Médecine de Fès. Les Annales de Médecine et de Thérapeutique 2009;1(1):4-10.
17. Brailovsky C, Charlin B, Beausoleil S, Cote S, Van der Vleuten C. Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: an experimental study on the script concordance test. Med Educ 2001;35(5):430-36.
18. Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CP. Script concordance testing: a review of published validity evidence. Med Educ 2011;45(4):329-38.
19. Groves M, O'Rourke P, Alexander H. The association between student characteristics and the development of clinical reasoning in a graduate-entry, PBL medical programme. Med Teach 2003;25(6):626-31.
20. Cilliers F, Schuwirth L, Herman N, Adendorff H, Van der Vleuten C. A model of the pre-assessment learning effects of summative assessment in medical education. Adv Health Sci Educ Theory Pract 2012 Mar;17(1):39-53.
21. Norcini J, Anderson B, Bollela V, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach 2011;33(3):206-14.