European Journal of
Dental Education

Evaluation of a new e-Learning resource for calibrating OSCE examiners on the use of rating scales

Journal:	*European Journal of Dental Education*
Manuscript ID	EJE-19-3291.R2
Manuscript Type:	Original Article
Keywords:	OSCE, rating scale, examiner training, calibration

Evaluation of a new e-learning resource for calibrating OSCE examiners on the use of rating scales.

Abstract:

Introduction: Rating scales have been described as better at assessing behaviours such as professionalism during Objective Structured Clinical Examinations (OSCEs). However, there is an increased need to train and calibrate staff on their use prior to student assessment.

Material and methods: An online e-learning package was developed and made available to all examiners at the Institute of Dentistry at the University of Aberdeen. The package included videos of three OSCE stations (medical emergency, rubber dam placement and handling a complaint) which were recorded in two different scenarios; (excellent and ~~borderline~~ unsatisfactory candidate). These videos were recorded to meet a predefined marking score. The examiners were required to mark the six videos using pre-set marking criteria (check list and rating scales). The rating scales included professionalism, general clinical ability and/or communication skills. For each video, examiners were given four possible options (unsatisfactory, borderline, satisfactory or excellent), they were provided with a description for each domain. They were also required to complete a questionnaire to gather their views on the use of this e-learning environment.

Results: Fifteen examiners completed the task. The total scores given were very similar to the expected scores for the medical emergency and complaint stations, however this was not the case for the rubber dam station (p-value 0.017 and 0.036). This could be attributed to some aspects of the placement of the rubber dam being unclear as commented on in the examiners questionnaires. There was consistency in the selection of marks on the rating scales (inter-examiner correlation ranged between 0.916 and 0.979).

Conclusion: Further studies are required on the field of e-learning training to calibrate examiners for practical assessment, however, this study provides preliminary evidence to support the use of videos as part of an online training package to calibrate OSCE examiners on the use of rating scales.

Introduction:

In the past two decades the focus on dental assessment has shifted to assessing clinical competence in an objective and structured fashion since Harden et al[1] described the use of Objective Structured Clinical Examinations (OSCEs) in medicine back in 1975. OSCEs were introduced in dentistry in the late 90s where Manogue et al[2] described the implementation of a dental OSCE on fourth year dental students with great validity and reliability. These exams have also been used to assess students in preclinical years, maintaining high standards of reliability[3]. Acceptability of this examination was greatly supported by staff[4] and students[2]. This examination allows examiners to assess students on specific and standardised tasks using multiple observations. Historically, those tasks had been assessed by means of using checklists which varied in the format of marking students, ranging from dichotomous options (yes/no) to multiple options of a specific task (done, partly done, attempted, not done) within a given station. This scoring system only allowed the examiner to observe the student's performance and identify which action they had performed, making the examiners into observers of behaviours rather than interpreters of behaviours[5].

Dreyfus and Dreyfus (1986)[6] observed that problem solving was addressed in different ways depending on the level of expertise of various professionals. They classified the five stages of development of expertise as novice, advanced beginner, competence, proficiency and expertise. Following from their research they concluded that experts cannot break down their thinking into small components and have difficulty in returning to the novice form of problem solving.

Therefore, several authors have suggested that checklists may penalize candidates that arrive at a diagnosis quickly and not following a specific checklist as a result of being at a more experienced level[5,7-9].

Instead, it has been proposed that the use of rating scales may be more valid than checklists[5] when formally assessing candidates. However, these are subject to interpretation and call for graded responses to a set of behaviours observed over a longer period of time[10]. For example, when an examiner is assessing a communication skills station, they will be evaluating all the content of that station, therefore, when it comes to using rating scales or global ratings (if using borderline regression method

for standard setting), the reliability will be higher as they have been able to observe and assess all the items throughout the station[10].

Rating scales consider skills performance across several domains using a Likert-type scale[5]; they are best used for assessing behaviours of candidates rather than the classical "done or not done" approach. They have been shown to discriminate better between students than the usual checklist once examiners had proper training on their use[5], and this training can be time consuming.

When it comes to training examiners, there are multiple difficulties encountered when trying to release them from their clinical commitments, and maybe e-learning tools might be a more realistic approach. E-learning can be effective and learner-centric, allowing users to decide when and where to learn[11].

The University of Aberdeen created a new online website named Assessment Central where all new policies and training in relation to assessment take place. Current available training is mainly focused on medical OSCE training and global ratings. The Institute of Dentistry in Aberdeen offers a graduate-entry 4-year Bachelors in Dental Surgery (BDS) course and limits the entrance to 20 students per year. As a result of this, we are not able to use global ratings for standard setting the OSCE papers by means of using a borderline regression method, instead we use the modified Angoff method. For this reason, we had to tailor the training to specific dental encounters and to train staff on the use of the newly developed rating scales.

This study aims to evaluate if a new e-learning resource can help calibrate examiners on the use of rating scales prior to a summative OSCE for undergraduate dental students. We also investigated the examiners' perception of the use of the resource and the perceived effectiveness.

Materials and methods:

Developing e-learning packages:

During a period of a month, we recorded six videos based on three dental OSCE stations previously used at the dental school (medical emergency, rubber dam placement and handling a complaint). These scenarios were chosen as each of them assessed different skills. The scenarios were recorded using staff members acting as patient (when required) and/or candidate. Each scenario was recorded twice; the first

video was recorded with a ~~borderline~~unsatisfactory candidate and the second with an excellent candidate.

Each scenario had a marking criteria that consisted of a combination of a checklist and rating scales. They all had a total score which was calculated by the assessment lead who was responsible for writing the scripts, recording and editing all the videos. The assessment lead had several years of experience in running and assessing this type of examination. This total score was the target value against which we also compared the results of the examiners that took place in this study. There was also a target mark for each of the rating scales used in each station, and this was compared to the ones given by each examiner.

After editing each video, these were uploaded to the Assessment Central website.

Each video lasted between three and five minutes, which is similar to the time that a student might take to complete the scenario in a real exam situation.

Examiner calibration exercise:

All possible examiners for the forthcoming dental OSCEs were contacted via e-mail and invited to participate in this study which involved marking the six videos using the pre-set marking criteria (a combination of a check list and rating scales). All the examiners had previous experience in assessing dental OSCES.

Examiners were also required to complete an anonymous, paper-based questionnaire at the end, to assess their perceptions of the usefulness of this new e-learning resource for calibrating examiners.

The questionnaire had demographic questions and questions relating to the usefulness of the videos. A summary of the questions can be found on table 1.

Examiners were able to do this at any time of the day, but it had to be done before the first day of the OSCEs. They were all given four weeks notice to complete the exercise. Examiners were only able to assess each video once. All the paperwork was collected by an administrator so that the results were blinded to the investigator.

The rating scales included up to three possible domains: professionalism, general clinical ability and/or communication skills. Each rating scale had four4 possible options (unsatisfactory, borderline, satisfactory or excellent) and for each of the rating

scales there was an anchor/grade descriptor attached to each option. The descriptors for these three rating scales can be found on table 2.

A short debrief on the preliminary data analysis was carried out prior to the start of the OSCE examinations, where rating scales were discussed again using examples of the videos that examiners had watched.

Statistics:

The results were collated in an SPSS file (IBM Corp. SPSS 24.0) which included an examiner number, the total score given for each of the six stations and the expected score. This score was calculated by the person responsible for recording and editing the videos and who was involved in writing the scripts for each video.

Statistical calculation included the mean and median scores across examiners for each station. T-test calculation was used to assess whether the mean score from the examiners was different from the expected score for each individual station. $P < 0.05$ was considered statistically significant. We also calculated the reliability of using the rating scales by means of an intra class correlation (ICC).

Ethics:

This study was approved by the Life Sciences and Medicine Ethics Review Board (CERB) at the University of Aberdeen.

Results:

Out of the twenty-five examiners invited to take part, 15 completed the task. The reason ten examiners did not complete the task was due to lack of time within the time frame given. The majority of examiners were females (n=10) and their ages ranged from 28 to 69 (mean 40.3 Standard deviation). Thirteen examiners were dentists, one a dental technician instructor (did not mark the rubber dam station) and another a dental nurse.

Table 3 provides a description of mean and median scores for each station. The mean score for medical emergency 2 and complaint 2 (excellent candidate) are similar to the expected score but that is not the case for rubber dam 1 and 2 and to some extent complaint 1 and medical emergency 1 (~~borderline~~ unsatisfactory candidate).

In the one sample t-test, we compared the mean of the 15 examiners, to the test value (expected score). Table 4 it shows the estimate of the difference from this test value and the 95% Confidence interval.

From this exercise we can report that there is no evidence of a difference between the examiners' mean score and the expected score for medical emergency 1 and 2 or complaint 2 (p=0.068, 0.166 and 0.403 respectively).

However, for rubber dam 1, there is a significant departure from the expected score (p=0.017) with a mean difference of 3.5 points lower on average by the examiners. The discrepancy for rubber dam 2 and complaint 1 is not quite as strong (p=0.036 and 0.013 respectively) but is still significant.

The reliability of the use of rating scales was calculated using the intra class correlation. This showed that there was consistency in the selection of marks using the rating scales. The Intra Cclass Ccorrelation (ICC) for the professionalism domain was 0.925, for general clinical ability 0.916 and for communication skills was 0.979.

Examiners perceptions:

All the examiners felt that the length of the videos (between three and five minutes) was appropriate and they all watched the videos on their university computer.

Comments in relation to the use of this resource were generally positive. Examples quotes include: "it gives the opportunity to re-observe procedures and do it at your own time" and "you fill in responses on your own without peer suggestion". They also commented on the type of stations used: "communication stations work very well on this type of resource" and "it gives good examples of good and bad candidates".

Although they also made some comments about the fact that some videos were not fit for this type of training: "possible camera adjustments required with more detailed OSCE stations such as the rubber dam one". One examiner also suggested that it would have been better to have a borderline passing students rather than excellent and failing student: "add video examples of borderline or satisfactory performances", however this did not affect the overall mark that examiners awarded.

Discussion:

Our results showed that the use of videos as part of an e-learning training package is a good way to calibrate staff on the use of rating scales as there was consistency in the selection of the right option for each rating scale used in these videos. Previous studies have also concluded that training programs using videos are necessary and effective in maintaining the integrity of high-stake assessments such as OSCEs[12], and videos have also been proven to be effective in the use of global rating scales[11,13]. Holmboe et al (2004)[14] also concluded that this type of training with direct observation produces changes in faculty evaluation behaviours and that this lasts for at least 8 months. Preusche et al (2012)[15] proposed twelve tips for designing and implementing a structured training for rating scales in OSCEs. In their proposal, they mention the need to use examples for the frame of reference and that videos are a very useful way of doing this when a common standard is not in existence. It is also mentioned how important videos can be to train the examiners in observation skills in order to use the rating scales appropriately.

However, Cooke et al as well as Byrne et al mentioned that no differences between trained and untrained assessors using video-taped resources as part of a face-to-face workshop were identified[16,17]. Byrne et al also reported that the training was perceived as highly effective, as in our study[17].

The majority of literature reviewed on the use of rating scales concentrates mainly on communication skills, professional behaviours and consultation skills[12,14-18]. This would explain why the scores that our examiners gave for the practical station (rubber dam) were so dissimilar. Some aspects of the rubber dam placement were not as clear on the video as others and as a result, examiners were not able to complete the checklist appropriately; this was also mentioned in their comments. In the future, practical stations, such as rubber dam placement, should be recorded with multiple cameras thus recording the task from different angles; this would allow examiners to view the station from different perspectives as they could do in a real exam situation.

When designing the rating scales, it was decided to have a scale of ~~4~~ four different options with specific descriptors to each option in each rating scale, as previously described in the literature[11]. That was based on the assumption that if five options were offered; the majority of examiners would pick the middle one when in doubt.

Some studies have suggested using a 5-point Likert scale with only points 1, 3 and 5 anchored to an explicit descriptor[5,7]. Chahine et al (2016)[19] used a scale of 1 (Inferior) to 6 (Excellent) when designing their 8 rating scales or competencies. However, when analysing their results they obtained very few scores in the Inferior and Excellent categories and combined them with adjacent scores for analysis, resulting in four categories that resemble the categories used in our study.

Our study had some limitations. The sample size was limited to the number of examiners available for the undergraduate BDS programme. The second limitation is that the videos focused on the two-end spectrum of possible candidates (excellent and ~~borderline/~~unsatisfactory) and did not investigate the full range of options within each rating scale. We believe that recordings from borderline candidates on these categories of rating scales would enrich and improve our training for examiners. Finally, the videos focused on only three OSCE stations, not allowing for more variety of marking schemes and grading scales. However, this is a pilot study, which will help us carry out more extensive studies on this field.

Conclusion:

Further studies are required on the field of e-learning training to calibrate examiners for practical assessment, however, this study provides preliminary evidence to support the use of videos as part of an online training package to calibrate OSCE examiners on the use of rating scales.

Conflict of interest: the authors have no conflict of interest to declare.

References:

1. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J 1975: 1: 447–451.
2. Manogue M, Brow G. Developing and implementing and OSCE in dentistry. Eur J Dent Educ 1998;2:51-57.

3. Eberhard L, Hassel A, Baumer A, Becker F, Beck-Mußotter J et al. Analysis of quality and feasibility of an objective structured clinical examination (OSCE) in preclinical dental education. Eur J Dent Educ 15 (2011) 172–178.
4. Schoonheim-Klein M, Walmsley AD, Habets L, van der Velden U and Manogue M. An implementation strategy for introducing an OSCE into a dental school. Eur J Dent Educ 2005; *9*: 143–149.
5. Regehr G, macRae H, Reznich RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad med 1998;73:993-997.
6. Dreyfus HL, Dreyfus SE. Mind over machine. New York: Free Press, 1986.
7. Hodges B, Regher G, McNaughton N, Tiberius R, Hanson M. OSCE checklist do not capture increasing levels of expertise. Acad med 1999;74:1129-1134.
8. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Medical Education 2003;37:1012-1016.
9. Hodges B, McNaughton N, Regehr G, Tiberius R, Hanson M. The challenges of creating new OSCE measurements to capture the characteristics of expertise. Medical Education 2002;36:742-748.
10. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Medical Education 2011;45:1181-1189.
11. Gormley GJ, Johnston J, Thomson C, Mcglade K. Awarding global grades in OSCEs: Evaluation of a novel eLearning resource for OSCE examiners. Medical teacher 2012;34:587-589.
12. Schwartzman E, Hsu DI, Law AV, Chung EP. Assessment of patient communication skills during OSCE: Examining effectiveness of a training program in minimizing inter-grader variability. Patient Education and Counseling 2011;8:472-477.
13. Reid K, Smallwood D, Collins M, Sutherland R, Dodds A. Taking OSCE examiner training on the road: reaching the masses. Med Educ Online 2016;21:32389.
14. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence. Ann intern Med 2004;140:874-881.

15. Preusche I, Schmidts M, Wagner-menghin M. Twelve tips for designing and implementing a structured rater training in OSCEs. Medical teacher 2012;34:368-372.
16. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of min-CEX scores: a randomized, controlled trial. J gen Intern med 2008;24(1):74-79.
17. Byrne A, Soskova T, Dawkins J, Coombes L. A pilot study of marking accuracy and mental workload as measures of OSCE examiner performance. BMC Medical Education 2016;16:191-196.
18. Davis R, Ellerton C, Evans C. Reaching consensus on measuring professional behaviours in physical therapy objective structures Clinical Examinations. Physiotherapy Canda 2017;69(1):65-72.
19. Chahine S, Holmes B, Kowsalewski Z. In the minds of OSCE examiners: uncovering hidden assumptions. Adv in Health Sci Educ Theory Pract. 2016;Aug;21(3):609-25.

Table 1: Questions asked on the questionnaire:

Questions	Answers
Did you get a clear three dimensional image of the procedure and its outcome?	Yes/No
Is the length of the videos enough to retain your concentration during the viewing?	Yes/No (possibility of free text)
Did you use the videos to check how the students are taught in the clinical skills laboratory?	Yes/No (possibility of free text)
Where did you watch the videos?	Laptop, home computer, university computer, smartphone, other
Do you think that video recorded procedures will improve the learning experience at Aberdeen Dental School?	Free text
How can the demonstration of the exercises be improved?	Free text

Table 2: Definition of the grade descriptors for the communication skills, general clinical ability and professionalism rating scales.

		Descriptors
Communication skills	Unsatisfactory	Lacks clarity or incoherent, essential information omitted, non-verbal communication could be interpreted as hostile; ignores patient's enquiries or does not respond adequately; blames others; judgemental; dismissive, does not listen, arrogant. Patient has grounds to complain
	Borderline	Answers and/or explanations require some clarification; patient may be unsure as to the tone of the communication. Candidate lacks a degree of confidence or appears ill at ease when communicating. Not engaged with patient.
	Satisfactory	Coherent answers and explanations, appropriate responses to patient's enquiries. Candidate is reasonably confident. Candidate has a friendly and caring disposition.
	Excellent	Confident, coherent and eloquent. Good listening skills. Sensitive to patient's non-verbal communication and able to respond accordingly. Complexity of explanation is matched to patient's level of understanding and interest. Very caring, friendly and empathetic attitude. Patient may be moved to write a testimonial.
General clinical ability	Unsatisfactory	Haphazard; the sequence of tasks performed by the candidate to execute the procedure had to be revised and/or repeated or was not attempted. Working environment was not ergonomic.
	Borderline	Poor organisation however still able to perform/attempt/approach the task without significant risk of injury or financial loss. An ergonomic working environment was attempted but not achieved.
	Satisfactory	Candidate's level of organisation was consistent with safe completion of task within the time frame. Good concession to ergonomic working environment however this was less than optimal.
	Excellent	Exceptional organisation; candidate performs procedure in an orderly and logical sequence consistent with an exemplary working knowledge and at all times maintains an ergonomic working environment.
Professionalism	Unsatisfactory	Offensive, fails to put patients' interest first, dishonest, rude, fails to take responsibility, judgemental, arrogant, lack of empathy or any other display of unprofessional conduct.

	Borderline	Inoffensive manner in that no clear evidence of poor professionalism however doesn't instil confidence.
	Satisfactory	Adequate professional manner and instils confidence in the patient.
	Excellent	Exemplary professional attitude, student instils confidence to the patient.

Table 3: Descriptive information for scores at each station

Station	**Expected Score**	N obs	Mean	SD	Median	IQR
Medical Emergency 1 (borderline candidate)	6	15	4.93	2.09	5	(4, 6)
Medical Emergency 2 (excellent candidate)	28	15	27.3	1.94	28	(26, 29)
Rubber dam (borderline candidate)	11	14	7.5	4.77	7	(3.75, 11.25)
Rubber dam (excellent candidate)	26	14	23.7	3.65	24	(20.5, 26.5)
Complaint (borderline candidate)	4	15	2.87	1.55	2	(2, 4)
Complaint (excellent candidate)	21	15	20.7	1.5	21	(20, 22)

N obs (number of observations), SD (Standard deviation), IQR (Interquartile range)

Table 4: One-sample t-test results

Situation	**Test value**	Mean diff	95% CI	p-value
Medical Emergency (borderline candidate)	6	-1.07	(-2.22, 0.09)	0.068
Medical Emergency (excellent candidate)	28	-0.73	(-1.81, 0.34)	0.166
Rubber dam (borderline candidate)	11	-3.50	(-6.25, -0.75)	0.017
Rubber dam (excellent candidate)	26	-2.29	(-4.39, -0.18)	0.036
Complaint (borderline candidate)	4	-1.13	(-1.99, -0.27)	0.013
Complaint (excellent candidate)	21	-0.33	(-1.16, 0.50)	0.403

Mean diff (mean difference), CI (Confidence Interval)

References:

1. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J 1975: 1: 447–451.
2. Manogue M, Brow G. Developing and implementing and OSCE in dentistry. Eur J Dent Educ 1998;2:51-57.
3. Eberhard L, Hassel A, Baumer A, Becker F, Beck-Mußotter J et al. Analysis of quality and feasibility of an objective structured clinical examination (OSCE) in preclinical dental education. Eur J Dent Educ 15 (2011) 172–178.
4. Schoonheim-Klein M, Walmsley AD, Habets L, van der Velden U and Manogue M. An implementation strategy for introducing an OSCE into a dental school. Eur J Dent Educ 2005; *9*: 143–149.
5. Regehr G, macRae H, Reznich RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad med 1998;73:993-997.
6. Dreyfus HL, Dreyfus SE. Mind over machine. New York: Free Press, 1986.
7. Hodges B, Regher G, McNaughton N, Tiberius R, Hanson M. OSCE checklist do not capture increasing levels of expertise. Acad med 199;74:1129-1134.
8. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Medical Education 2003;37:1012-1016.
9. Hodges B, McNaughton N, Regehr G, Tiberius R, Hanson M. The challenges of creating new OSCE measurements to capture the characteristics of expertise. Medical Education 2002;36:742-748.
10. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Medical Education 2011;45:1181-1189.
11. Gormley GJ, Johnston J, Thomson C, Mcglade K. Awarding global grades in OSCEs: Evaluation of a novel eLearning resource for OSCE examiners. Medical teacher 2012;34:587-589.
12. Schwartzman E, Hsu DI, Law AV, Chung EP. Assessment of patient communication skills during OSCE: Examining effectiveness of a training program in minimizing inter-grader variability. Patient Education and Counseling 2011;8:472-477.

13. Reid K, Smallwood D, Collins M, Sutherland R, Dodds A. Taking OSCE examiner training on the road: reaching the masses. Med Educ Online 2016;21:32389.
14. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence. Ann intern Med 2004;140:874-881.
15. Preusche I, Schmidts M, Wagner-menghin M. Twelve tips for designing and implementing a structured rater training in OSCEs. Medical teacher 2012;34:368-372.
16. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of min-CEX scores: a randomized, controlled trial. J gen Intern med 2008;24(1):74-79.
17. Byrne A, Soskova T, Dawkins J, Coombes L. A pilot study of marking accuracy and mental workload as measures of OSCE examiner performance. BMC Medical Education 2016;16:191-196.
18. Davis R, Ellerton C, Evans C. Reaching consensus on measuring professional behaviours in physical therapy objective structures Clinical Examinations. Physiotherapy Canda 2017;69(1):65-72.
19. Chahine S, Holmes B, Kowsalewski Z. In the minds of OSCE examiners: uncovering hidden assumptions. Adv in Health Sci Educ 2016;21:609-625.

Evaluation of a new e-learning resource for calibrating OSCE examiners on the use of rating scales.

Abstract:

Introduction: Rating scales have been described as better at assessing behaviours such as professionalism during Objective Structured Clinical Examinations (OSCEs). However, there is an increased need to train and calibrate staff on their use prior to student assessment.

Material and methods: An online e-learning package was developed and made available to all examiners at the Institute of Dentistry at the University of Aberdeen. The package included videos of three OSCE stations (medical emergency, rubber dam placement and handling a complaint) which were recorded in two different scenarios; (excellent and unsatisfactory candidate). These videos were recorded to meet a predefined marking score. The examiners were required to mark the six videos using pre-set marking criteria (check list and rating scales). The rating scales included professionalism, general clinical ability and/or communication skills. For each video, examiners were given four possible options (unsatisfactory, borderline, satisfactory or excellent), they were provided with a description for each domain. They were also required to complete a questionnaire to gather their views on the use of this e-learning environment.

Results: Fifteen examiners completed the task. The total scores given were very similar to the expected scores for the medical emergency and complaint stations, however this was not the case for the rubber dam station (p-value 0.017 and 0.036). This could be attributed to some aspects of the placement of the rubber dam being unclear as commented on in the examiners questionnaires. There was consistency in the selection of marks on the rating scales (inter-examiner correlation ranged between 0.916 and 0.979).

Conclusion: Further studies are required on the field of e-learning training to calibrate examiners for practical assessment, however, this study provides preliminary evidence to support the use of videos as part of an online training package to calibrate OSCE examiners on the use of rating scales.

Introduction:

In the past two decades the focus on dental assessment has shifted to assessing clinical competence in an objective and structured fashion since Harden et al[1] described the use of Objective Structured Clinical Examinations (OSCEs) in medicine back in 1975. OSCEs were introduced in dentistry in the late 90s where Manogue et al[2] described the implementation of a dental OSCE on fourth year dental students with great validity and reliability. These exams have also been used to assess students in preclinical years, maintaining high standards of reliability[3]. Acceptability of this examination was greatly supported by staff[4] and students[2]. This examination allows examiners to assess students on specific and standardised tasks using multiple observations. Historically, those tasks had been assessed by means of using checklists which varied in the format of marking students, ranging from dichotomous options (yes/no) to multiple options of a specific task (done, partly done, attempted, not done) within a given station. This scoring system only allowed the examiner to observe the student's performance and identify which action they had performed, making the examiners into observers of behaviours rather than interpreters of behaviours[5].

Dreyfus and Dreyfus (1986)[6] observed that problem solving was addressed in different ways depending on the level of expertise of various professionals. They classified the five stages of development of expertise as novice, advanced beginner, competence, proficiency and expertise. Following from their research they concluded that experts cannot break down their thinking into small components and have difficulty in returning to the novice form of problem solving.

Therefore, several authors have suggested that checklists may penalize candidates that arrive at a diagnosis quickly and not following a specific checklist as a result of being at a more experienced level[5,7-9].

Instead, it has been proposed that the use of rating scales may be more valid than checklists[5] when formally assessing candidates. However, these are subject to interpretation and call for graded responses to a set of behaviours observed over a longer period of time[10]. For example, when an examiner is assessing a communication skills station, they will be evaluating all the content of that station, therefore, when it comes to using rating scales or global ratings (if using borderline regression method

for standard setting), the reliability will be higher as they have been able to observe and assess all the items throughout the station[10].

Rating scales consider skills performance across several domains using a Likert-type scale[5]; they are best used for assessing behaviours of candidates rather than the classical "done or not done" approach. They have been shown to discriminate better between students than the usual checklist once examiners had proper training on their use[5], and this training can be time consuming.

When it comes to training examiners, there are multiple difficulties encountered when trying to release them from their clinical commitments, and maybe e-learning tools might be a more realistic approach. E-learning can be effective and learner-centric, allowing users to decide when and where to learn[11].

The University of Aberdeen created a new online website named Assessment Central where all new policies and training in relation to assessment take place. Current available training is mainly focused on medical OSCE training and global ratings. The Institute of Dentistry in Aberdeen offers a graduate-entry 4-year Bachelors in Dental Surgery (BDS) course and limits the entrance to 20 students per year. As a result of this, we are not able to use global ratings for standard setting the OSCE papers by means of using a borderline regression method, instead we use the modified Angoff method. For this reason, we had to tailor the training to specific dental encounters and to train staff on the use of the newly developed rating scales.

This study aims to evaluate if a new e-learning resource can help calibrate examiners on the use of rating scales prior to a summative OSCE for undergraduate dental students. We also investigated the examiners' perception of the use of the resource and the perceived effectiveness.

Materials and methods:

Developing e-learning packages:

During a period of a month, we recorded six videos based on three dental OSCE stations previously used at the dental school (medical emergency, rubber dam placement and handling a complaint). These scenarios were chosen as each of them assessed different skills. The scenarios were recorded using staff members acting as patient (when required) and/or candidate. Each scenario was recorded twice; the first

video was recorded with aunsatisfactory candidate and the second with an excellent candidate.

Each scenario had a marking criteria that consisted of a combination of a checklist and rating scales. They all had a total score which was calculated by the assessment lead who was responsible for writing the scripts, recording and editing all the videos. The assessment lead had several years of experience in running and assessing this type of examination. This total score was the target value against which we also compared the results of the examiners that took place in this study. There was also a target mark for each of the rating scales used in each station, and this was compared to the ones given by each examiner.

After editing each video, these were uploaded to the Assessment Central website.

Each video lasted between three and five minutes, which is similar to the time that a student might take to complete the scenario in a real exam situation.

Examiner calibration exercise:

All possible examiners for the forthcoming dental OSCEs were contacted via e-mail and invited to participate in this study which involved marking the six videos using the pre-set marking criteria (a combination of a check list and rating scales). All the examiners had previous experience in assessing dental OSCES.

Examiners were also required to complete an anonymous, paper-based questionnaire at the end, to assess their perceptions of the usefulness of this new e-learning resource for calibrating examiners.

The questionnaire had demographic questions and questions relating to the usefulness of the videos. A summary of the questions can be found on table 1.

Examiners were able to do this at any time of the day, but it had to be done before the first day of the OSCEs. They were all given four weeks notice to complete the exercise. Examiners were only able to assess each video once. All the paperwork was collected by an administrator so that the results were blinded to the investigator.

The rating scales included up to three possible domains: professionalism, general clinical ability and/or communication skills. Each rating scale had four possible options (unsatisfactory, borderline, satisfactory or excellent) and for each of the rating scales

there was an anchor/grade descriptor attached to each option. The descriptors for these three rating scales can be found on table 2.

A short debrief on the preliminary data analysis was carried out prior to the start of the OSCE examinations, where rating scales were discussed again using examples of the videos that examiners had watched.

Statistics:

The results were collated in an SPSS file (IBM Corp. SPSS 24.0) which included an examiner number, the total score given for each of the six stations and the expected score. This score was calculated by the person responsible for recording and editing the videos and who was involved in writing the scripts for each video.

Statistical calculation included the mean and median scores across examiners for each station. T-test calculation was used to assess whether the mean score from the examiners was different from the expected score for each individual station. $P < 0.05$ was considered statistically significant. We also calculated the reliability of using the rating scales by means of an intra class correlation (ICC).

Ethics:

This study was approved by the Life Sciences and Medicine Ethics Review Board (CERB) at the University of Aberdeen.

Results:

Out of the twenty-five examiners invited to take part, 15 completed the task. The reason ten examiners did not complete the task was due to lack of time within the time frame given. The majority of examiners were females (n=10) and their ages ranged from 28 to 69 (mean 40.3 Standard deviation). Thirteen examiners were dentists, one a dental technician instructor (did not mark the rubber dam station) and another a dental nurse.

Table 3 provides a description of mean and median scores for each station. The mean score for medical emergency 2 and complaint 2 (excellent candidate) are similar to the expected score but that is not the case for rubber dam 1 and 2 and to some extent complaint 1 and medical emergency 1 (unsatisfactory candidate).

In the one sample t-test, we compared the mean of the 15 examiners, to the test value (expected score). Table 4 it shows the estimate of the difference from this test value and the 95% Confidence interval.

From this exercise we can report that there is no evidence of a difference between the examiners' mean score and the expected score for medical emergency 1 and 2 or complaint 2 (p=0.068, 0.166 and 0.403 respectively).

However, for rubber dam 1, there is a significant departure from the expected score (p=0.017) with a mean difference of 3.5 points lower on average by the examiners. The discrepancy for rubber dam 2 and complaint 1 is not quite as strong (p=0.036 and 0.013 respectively) but is still significant.

The reliability of the use of rating scales was calculated using the intra class correlation. This showed that there was consistency in the selection of marks using the rating scales. The Intra Class Correlation (ICC) for the professionalism domain was 0.925, for general clinical ability 0.916 and for communication skills was 0.979.

Examiners perceptions:

All the examiners felt that the length of the videos (between three and five minutes) was appropriate and they all watched the videos on their university computer.

Comments in relation to the use of this resource were generally positive. Examples quotes include: "it gives the opportunity to re-observe procedures and do it at your own time" and "you fill in responses on your own without peer suggestion". They also commented on the type of stations used: "communication stations work very well on this type of resource" and "it gives good examples of good and bad candidates".

Although they also made some comments about the fact that some videos were not fit for this type of training: "possible camera adjustments required with more detailed OSCE stations such as the rubber dam one". One examiner also suggested that it would have been better to have a borderline passing students rather than excellent and failing student: "add video examples of borderline or satisfactory performances", however this did not affect the overall mark that examiners awarded.

Discussion:

Our results showed that the use of videos as part of an e-learning training package is a good way to calibrate staff on the use of rating scales as there was consistency in the selection of the right option for each rating scale used in these videos. Previous studies have also concluded that training programs using videos are necessary and effective in maintaining the integrity of high-stake assessments such as OSCEs[12], and videos have also been proven to be effective in the use of global rating scales[11,13]. Holmboe et al (2004)[14] also concluded that this type of training with direct observation produces changes in faculty evaluation behaviours and that this lasts for at least 8 months. Preusche et al (2012)[15] proposed twelve tips for designing and implementing a structured training for rating scales in OSCEs. In their proposal, they mention the need to use examples for the frame of reference and that videos are a very useful way of doing this when a common standard is not in existence. It is also mentioned how important videos can be to train the examiners in observation skills in order to use the rating scales appropriately.

However, Cooke et al as well as Byrne et al mentioned that no differences between trained and untrained assessors using video-taped resources as part of a face-to-face workshop were identified[16,17]. Byrne et al also reported that the training was perceived as highly effective, as in our study[17].

The majority of literature reviewed on the use of rating scales concentrates mainly on communication skills, professional behaviours and consultation skills[12,14-18]. This would explain why the scores that our examiners gave for the practical station (rubber dam) were so dissimilar. Some aspects of the rubber dam placement were not as clear on the video as others and as a result, examiners were not able to complete the checklist appropriately; this was also mentioned in their comments. In the future, practical stations, such as rubber dam placement, should be recorded with multiple cameras thus recording the task from different angles; this would allow examiners to view the station from different perspectives as they could do in a real exam situation.

When designing the rating scales, it was decided to have a scale of four different options with specific descriptors to each option in each rating scale, as previously described in the literature[11]. That was based on the assumption that if five options were offered; the majority of examiners would pick the middle one when in doubt.

Some studies have suggested using a 5-point Likert scale with only points 1, 3 and 5 anchored to an explicit descriptor[5,7]. Chahine et al (2016)[19] used a scale of 1 (Inferior) to 6 (Excellent) when designing their 8 rating scales or competencies. However, when analysing their results they obtained very few scores in the Inferior and Excellent categories and combined them with adjacent scores for analysis, resulting in four categories that resemble the categories used in our study.

Our study had some limitations. The sample size was limited to the number of examiners available for the undergraduate BDS programme. The second limitation is that the videos focused on the two-end spectrum of possible candidates (excellent and unsatisfactory) and did not investigate the full range of options within each rating scale. We believe that recordings from borderline candidates on these categories of rating scales would enrich and improve our training for examiners. Finally, the videos focused on only three OSCE stations, not allowing for more variety of marking schemes and grading scales. However, this is a pilot study, which will help us carry out more extensive studies on this field.

Conclusion:

Further studies are required on the field of e-learning training to calibrate examiners for practical assessment, however, this study provides preliminary evidence to support the use of videos as part of an online training package to calibrate OSCE examiners on the use of rating scales.

Conflict of interest: the authors have no conflict of interest to declare.

References:

1. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J 1975: 1: 447–451.
2. Manogue M, Brow G. Developing and implementing and OSCE in dentistry. Eur J Dent Educ 1998;2:51-57.

3. Eberhard L, Hassel A, Baumer A, Becker F, Beck-Mußotter J et al. Analysis of quality and feasibility of an objective structured clinical examination (OSCE) in preclinical dental education. Eur J Dent Educ 15 (2011) 172–178.
4. Schoonheim-Klein M, Walmsley AD, Habets L, van der Velden U and Manogue M. An implementation strategy for introducing an OSCE into a dental school. Eur J Dent Educ 2005; *9*: 143–149.
5. Regehr G, macRae H, Reznich RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad med 1998;73:993-997.
6. Dreyfus HL, Dreyfus SE. Mind over machine. New York: Free Press, 1986.
7. Hodges B, Regher G, McNaughton N, Tiberius R, Hanson M. OSCE checklist do not capture increasing levels of expertise. Acad med 1999;74:1129-1134.
8. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Medical Education 2003;37:1012-1016.
9. Hodges B, McNaughton N, Regehr G, Tiberius R, Hanson M. The challenges of creating new OSCE measurements to capture the characteristics of expertise. Medical Education 2002;36:742-748.
10. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Medical Education 2011;45:1181-1189.
11. Gormley GJ, Johnston J, Thomson C, Mcglade K. Awarding global grades in OSCEs: Evaluation of a novel eLearning resource for OSCE examiners. Medical teacher 2012;34:587-589.
12. Schwartzman E, Hsu DI, Law AV, Chung EP. Assessment of patient communication skills during OSCE: Examining effectiveness of a training program in minimizing inter-grader variability. Patient Education and Counseling 2011;8:472-477.
13. Reid K, Smallwood D, Collins M, Sutherland R, Dodds A. Taking OSCE examiner training on the road: reaching the masses. Med Educ Online 2016;21:32389.
14. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence. Ann intern Med 2004;140:874-881.

15. Preusche I, Schmidts M, Wagner-menghin M. Twelve tips for designing and implementing a structured rater training in OSCEs. Medical teacher 2012;34:368-372.
16. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of min-CEX scores: a randomized, controlled trial. J gen Intern med 2008;24(1):74-79.
17. Byrne A, Soskova T, Dawkins J, Coombes L. A pilot study of marking accuracy and mental workload as measures of OSCE examiner performance. BMC Medical Education 2016;16:191-196.
18. Davis R, Ellerton C, Evans C. Reaching consensus on measuring professional behaviours in physical therapy objective structures Clinical Examinations. Physiotherapy Canda 2017;69(1):65-72.
19. Chahine S, Holmes B, Kowsalewski Z. In the minds of OSCE examiners: uncovering hidden assumptions. Adv in Health Sci Educ Theory Pract. 2016;Aug;21(3):609-25.