Original Article

The Dilemma of Standard Setting for the OSCE

M Y Sukkar*

ABSTRACT

Background: Recently disparities between the OSCE raw scores and global scores have resulted in the need treatment of the raw scores in different ways such as borderline regression and borderline group regression. The object of this paper is to present station scoring forms designed to satisfy predetermined criteria and minimum pass levels.

Methods: Available samples of marking sheets and checklists designed by various examining bodies were scrutinized. Criteria were prioritized according to commonly used grading systems. The *station rating scale (check list)* was designed to allow the observer to concentrate on checking the performance of the candidate without marking at the same time. The *station marks form* enables entry of marks based on the criteria in the *Station rating scale.*

Results: Three forms were designed. Forms 1 & 2 should be prepared beforehand with real or standardized patients. Form 3 is a combination to be used when last minute stations are introduced. The mark allocated to each observable criterion is made within the limits of the specified criteria. The global score has been retained to check for inconsistencies and for longitudinal studies on validity and reliability.

Conclusions: Prototype forms are presented; using predetermined, categorized grading criteria. The forms enable examiners to separate the observation stage from the actual allocation of marks. As in all OSCE settings, objectivity, validity and reliability will depend on prioritizing the selection of stations, clarity of the selected criteria and the training of examiners.

Keywords:

"Criterion referenced" standard setting and minimum pass levels (MPL), have been widely agreed to represent the best yardstick against which pass/fail decisions and grading of students can be made[1]. However, many institutions resort to norm referencing and some wellknown authorities advocate "borderline regression" based on a global score decided by the examiner, or recently "borderline group" regression despite the claimed objectivity of the OSCE[2.3].

Although the OSCE psychometric characteristicshave been examined by numerous publications, the challenge of setting standards still exists[4].

*Professor of physiology, Educ. Development Unit, Nile College, Sudan.
Email: profmys@gmail.com

Marking schemes adopted by many institutions mark students within a narrow range defined as clear pass, border line and clear fail or to use open ended marking checklists to allot marks for each station. Allocating marks for the above three and sometimes four grades results in over marking and tends to lump students within a narrow range of marks which precludes the desirable discrimination[5].

The OSCE is now widely used by medical schools in the final MBBS and has to a large extent replaced the classical long case, short cases and clinical oral examinations[6]. Hence the need for a marking system for OSCE stations which is objective, reliable and not least, compatible with the other parts of the examination. That is to say the grades obtained from the station forms and the

overall assessment are compatible with those adopted by examination regulations which allocates percentages for the grades of A, B+, B, C, and F in most medical schools. If the station marking forms take this consideration into account in the checklists performance criteria, it will obviate resorting to "norm referencing" or rescaling of the raw scores.

Another concern is the tendency to over mark or under mark; leading to lack of congruence between the global score and the raw score and consequently the need for norm referencing or worse still adding raw marks .

The purpose of this paper is to present a scheme of marking and grading of the OSCE taking the above concerns into account hopefully approaching criterion referencing as far as possible.

MATERIALS AND METHODS:

Several samples of marking sheets used by various examining bodies were scrutinized so as to design new OSCE station forms. It was found that there are many examples of station marking sheets some of which reflect the drawbacks stated above[7,8].

Station rating scale

The first step in the design of the *Station rating scale is* to take care of specifying criteria for the observation of examiners. To do this the criteria are defined as essential, important, and other more advanced ones which can be used for grades of excellent or very good. These three categories should be spelled out by the panel of examiners in as unequivocal terms as possible. The *station rating scale (check list)* is intended to allow the observer to concentrate on his task of assessing the performance of the candidate without busying himself by allocating marks or grades at the same time. The results of the rating scale are followed by the allocation of marks after completion of the station. This division of the examiners' task should result in more accurate marking, a less subjective global score, less interexaminer variation and increase the objectivity and reliability of the instrument.

Each *station rating scale* should give opportunities for assessment of the "process" of carrying out the task and the "content" of the task itself (be it a communication component a specific case history or clinical examination). It goes without saying that the "process" deserves more weight in the criteria of the attitude and communication domain, while the checklists for other stations will give the process less weight than the actual task of obtaining a good history and performing a good clinical examination and reporting of findings, interpretation and management plan. The rating scale allows for a five point grading system which matches that used in other parts of the examination.

Spelling out the criteria for each station is a crucial step and no effort should be spared by the experts in their specification as these will determine the allocation of marks and therefore the grade decisions. Space is provided for remarks by the examiner to justify a fail. Such remarks will come in useful when considering marginal fail cases in examiners meetings. For the sake of validity, the number of stations in the various domains should receive special attention prior to and during the preparation phase.

The station marks form

The second consideration is to make sure that the pass level cannot be obtained by compensation within each *station marks form* or the total obtained by adding the marks of all the stations. The *station marks form* enables the examiners assessment entered in the station rating scale to be converted into actual marks based on the criteria and their categorization in the *Station rating scale.*

The *station marks form* is divided into three sections representing the main domains measured by the OSCE stations ie

attitude and communication skills, history taking, the clinical examination, interpretation and management plan. This is to ensure balanced assessment of the essential clinical competencies in each station; especially in a high stakes examination designed for a final MBBS examination. Each domain of the check list is therefore divided into categories of criteria; essential, important and others. The last category is not essential for pass but is important for grading. Defining and listing the categories should be done by a panel of experienced examiners.

Combined station rating scale and marks form

Taking into consideration the difficulty of availability of patients, last minute inclusion of stations is a common occurrence in some medical schools. This will make it impractical to use the station forms above. For such scenarios a third alternative form has been designed to accommodate such eventualities.

Training of examiners

Last, but by no means least, is the selection and familiarizing of examiners with the criteria for the station and the use of the rating scales as well as the conduction of the observation/ interpretation/explanation of the management plan. A structured induction meeting is recommended to be mandatory prior to the examination[9].

RESULTS:

Three forms were designed. Forms 1 and 2 should be prepared beforehand with standardized or simulated patients and checked by a panel. Form 3 can be used when last minute stations are introduced due to a variety of reasons.

The design of the marks form takes the concerns of standard setting criteria and alignment of the OSCE instrument with the institutional grading system into account. Form (1) shows a proposed prototype station rating scale/check list formatted to record the observations of the examiner without distracting his attention by allocation of marks and having to decide on pass/fail or grading of the candidate at the same time. This form can be modified to suit the objectives tested at a particular station.

Each station content will have a scale of categorized criteria (Form (2)) drawn according to the weight of competencies (i.e. essential, important, and other) required for the particular station case content area. The result obtained in each *station rating scale* is then transferred to the station marks form thus converting them into actual marks compatible with the institution's grading system. Hence the use of a marking scheme in which the examiner awards actual marks based on categorized criteriaand a rating scale identifying the level of performance.

An alternative (Form (3)) is presented as an example of the flexibility of this form to suit the objectives tested at any specific station. Form 3 can be modified to be relevant to the emphasis in each station; as long as the structure of the three sections is preserved to conform to the alignment between the school's grading system and the marks allocated to the OSCE part of the whole examination and that the pass mark cannot be obtained by compensation within the station marking scheme.

The mark allocated to each observable criterion enables examiners to make judgments within the limits of the criteria specified for each station (criterion reference). However, the global score has been retained in case of inconsistencies between stations, which can then be discussed in the examiners meeting. The global score can also serve the purpose of carrying out longitudinal studies on validity and reliability of the new forms and comparison to other studies.

Form (3) is an alternative designed to meet the eventuality of a station identified at the last minute for a legitimate reason.

Form No.1: OSCE station rating scale
Dept……. ……. Station No……
Student's ID……. Date……
To be filled by station examiner: *please tick (√) the appropriate box*

Communication						
OBSERVATION CRITERIA	Excellent	V. good	Good	Accept-able	Poor or Not done	Remarks
ESSENTIAL CRITERIA						
IMPORTANT CRITERIA						
Other criteria						
History						
ESSENTIAL CRITERIA						
IMPORTANT CRITERIA						
Other criteria						
Physical examinationl/ clinical findings						
ESSENTIAL CRITERIA						
IMPORTANT CRITERIA						
Other criteria						
Clinical reasoning & management						
ESSENTIAL CRITERIA						
IMPORTANT CRITERIA						
Other criteria						

After filling this form, transfer the results to marks in form No. 2

Form No. 2: Marks Form for OSCE Station
Dept.:……. Station No…… Student's ID…… Date……
Instructions to examiner: *Please write the most appropriate score according to performance criteria provided in the station marking sheet.*

Grade	***Excellent 8 or 9***	***V. good 7***	***Good 6***	***Pass 5***	***Clear Fail 1 or 2***	***Remarks***
Level of performance for grading	*• All criteria satisfied • Done well*	*• Most criteria satisfied • Done well*	*• A significant no. of criteria satisfied • Done reasonably well*	*• Essential Criteria done •Acceptable performance*	*Major Deficits*	
Communication : Initial stage of Dialogue or encounter	(according to station content)	*• Most criteria Satisfied • Done well*	*• A significant No. of criteria • Done reasonably well*	*• Essential Criteria done •Acceptable performance*	*Major faults*	
Score						Remarks
History, Body of Dialogue or skill	(according to station content)	*• Most criteria Satisfied • Done well*	*• A significant No. of criteria • Done reasonably well*	*• Essential Criteria done •Acceptable performance*	*Major deficiencies or inaccuracies*	
Score						
Clinical exam. Results / Interpretation / management plan	Correctness of information received & given (according to examiner's form)	*• Most criteria Satisfied • Done well*	*• A significant No. of criteria • Done reasonably well*	*• Essential Criteria done •Acceptable performance*	*Major deficiencies or inaccuracies*	
Sub Total						
Max. and Min. score	24-27				< 15	
Pass mark: 15 *Global score:*	**Total grade** (Circle one Grade) A B+ B C F		Examiner's Name:_____ Signature _____			

This form can be 'calibrated' by at least two examiners when such stations are included on the day of the examination. This form combines a criterion referenced marking scheme with the marking sheet and also provides room for the global score. However, it lacks the deliberate identification of criteria by a panel of examiners

DISCUSSION:

There have been many concerns about the validity and reliability of the OSCE since its inception by Harden in 1975[10]. These concerns have to a large extent been addressed through modification of check lists, training of examiners and simulated patients. Eventually, there seems to be agreement about setting a pass level by using the borderline regression method which is used to rescale the raw scores to fit a global assessment by expert examiners[11,12]. No doubt this process introduces an element of subjectivity and a

great deal of inter-examiner variation. It appears as if what matters is the overall assessment or "global score" identifying borderline candidates as seen by the examiner; no matter how much effort has been put in the examination[13]. The proposed marking scheme addresses alignment between the raw score which is criterion referenced and the grading system adopted by each institution.

Form No. 3: Criterion referenced station rating scale & marks sheet
Station No…… Students ID…… Date ……..
Instructions to examiner: *Please tick (√) the appropriate criterion box &circle the most appropriate score according to performance criteria rating scale at the end of the session.* A = excellent: *All criteria done well;* B+= Very good*: Most criteria done well;* B= Good: *Essential criteria done reasonably well;* C= Pass: *Essential criteria done with acceptable performance;* F= Fail: *Major deficiencies or omissions in essential (must do) criteria.*
List of criteria

	Must do	**A**	**B+**	**B**	**C**	**F**	**Remarks**
1.							
2.							
3.							
4.							
5.							
	Other criteria						
6.							
7.							
8.							
9.							
10.							
Circle one score (1 to 9)		**9**	**7.5**	**6.5**	**5.5**	**4 , 3**	**Global score: A B+ B C F**
		8	**7.0**	**6.0**	**5.0**	**2 , 1**	

Examiner'sname…….signature……

Form (1) shows that by checking the appropriate mark, the examiner aligns his marks to the system adopted by the institution; (e.g. 50 = Pass (C), 60 = Good (B), 70 = V. good (B+) and 80 = Excellent (A). As the allocation of marks in the *station marks form* has been designed to avoid or minimize compensation of failure in significant competencies by other less important parts of the station content areas, one would expect results which are criterion referenced; as well as compatible with the adopted grading scheme. Such scores can be safely added to obtain an overall grade in a particular discipline for the student's transcript.

Using the above marking scheme students will fall within the grade criteria spelled out in the examination regulation (e.g. A ≥ 80%, B^+ = 70-79, B = 60-69, C= 50-59 & F <50).

As the main purpose of the design of these forms is to increase its reliability and discrimination, the following examples illustrate this point:

1- A student scoring a pass will receive 5x3=15 out of a maximum of 30 , his mark will be 50%
2- A student who scores a clear fail in two

3- domains + excellent in one cannot compensate (i.e. 2+2+9=13) his mark will be <50%
4- A student who scores a clear fail in one domain can only compensate if he is more than just pass in the other two (e.g. 2+6+7=15)

There is no shortage of check lists for specific stations. We only need to train our examiners to identify the checklist criteria as essential, important and others so as to enable the conversion of the grades to actual marks.The presented scheme tests the domains of communication, clinical skills and clinical reasoning to ensure clinical content validity. This requires blue printing in the selection of station content and decisions on the appropriate form structure and content. A score of at least 50% should be obtained when all station scores are added. However, the student must pass aspecified number of stations (e.g. at least two thirds or more of the stations) to safeguard against compensation between stations in the presence of major deficiencies.

Professional attitude and communication skills are tested in most clinical encounter stations, after making sure that observers at these stations are aware of communication skills parameters. In addition, the blue print should include at least one station where communication skills and professional attitudes are the main content area.

CONCLUSION:

To achieve criterion referenced standard setting for the OSCE, these prototype forms use categorized grading criteria selected by a panel of experts. Several options of rating scales and marks forms are presented so as to meet the needs of different settings. The forms enable examiners to separate the observation stage from the actual allocation of marks. As in all OSCE settings, objectivity, precision and reliability will depend on prioritizing the selection of stations, clarity of the selected criteria and the training of examiners.

REFERENCES:

1. Sara MortazHejri, Mohammed Jalili, Arno M. Muijtien and Cees P.M. Van Der Vieutem. Assessing the reliability of borderline regressions method as a standard setting procedure for objective structured clinical examination, J Res Med Sc 2013, 18(10): 887 - 891.
2. Cohen, R, Rothman, A I, Poldre, P, Ross, J. Validity and gneralizability of global ratings in an objective structured clinical examination. Acad Med ,1991; 66:545-8
3. Wilkinson, T J Newble, D I, Frampton, C M. Standard setting in an objective structured clinical examination use of global ratings and borderline performance to determine the passing score. Med Educ.2001, 35:1043 – 1049.
4. Norcimi, J.J, Setting standards on educational tests. Med.Educ.2005;37: 464 -469
5. Pell, R, Fuller, R, Homer, M, Roberts T. International association for medical Education. How to measure quality of the OSCE: A review Adv Health Sc: Educ of metrics – AMEE guide no.49. Med Teach. 2010; 32: 802-11.
6. Newble ,D, Techniques for measuring clinical competence: Objective structured clinical examination. Med Educ 2004;38:199-203.
7. OCSE station forms Neel Burton, 2009/ Clinical skills for OSCEs,3rd edition [ISBN 9781904842590]Published by Scion Publishing Ltd.
8. Assessment Resource Center: Types of Assessment methods. OSCE rubrics. http://ar.cetl.hku/om_osce.htm
9. Newble D I , Hoare J , Sheldrake PF. The selection and training and examiners for clinical examinations. Med Educ 1980,19: 345-349.
10. Harden, R M, Stevenson, M,Downie W W, Wilson,G M. Assessment of clinical competence usingobjective structured examinations.BMJ 1975, 1: 447-51.
11. Reid, K J, Dodds, A. Comparing borderline group and borderline regression approaches to setting OSCE cut scores. J. Contemp. Med educ. 2014;2: 8-12.
12. Gormly, G, Summative OSCE in undergraduate medical education, Ulster med J 2011, 80: 127 - 132
13. Pell, G, Homer, M S ,and Fuller, R, Investigating disparity between global grades and check list scores in OSCEs. Medical teacher, 2015 (in press), http: //www.ncbi.nih.gov/pubmed/25683174.