Reliability of Watched Objective Structured Clinical Examination (WOSCE) Scores in Family Medicine Clerkship

Prof Faisal A. Latif Alnasir FPC, FRCGP, MICGP, PhD
Professor of Family Medicine
Vice President
Arabian Gulf University

Dr Qassim shboul PhD
Associate professor; Biostatistics
College Of Medicine and Medical Sciences
Arabian Gulf University

Prof Hossam Hamdy* FRCS
Professor of Surgery
Dean
College Of Medicine and Medical Sciences
Arabian Gulf University

*****Former Dean

Key words: WOSCE, family medicine rotation, scores, validity, inter-rater reliability.

Short running title: Reliability of WOSCE Examination

Corresponding Author: Prof Faisal A. Latif Alnasir
Vice President
Arabian Gulf University
Tel +973 17239701
Fax +973 17273456
P.O. Box 22979, Manama, Kingdom of Bahrain
E-mail: faisal.alnasir@gmail.com

Reliability of Watched Objective Structured Clinical Examination (WOSCE) Scores in Family Medicine Clerkship

Abstract

Background: At the College of Medicine and Medical Sciences of the Arabian Gulf University a new exam was implemented to examine year 6 students at the end of their family medicine clerkship rotation. It is called the watched objective structured clinical examination (WOSCE). The WOSCE was tried for four successive years since its development.

Objective: To study the interrater-reliability between the WOSCE's examiners.

Methods: The WOSCE consisted of 11 stations testing various areas of clinical competencies of year 6 medical students.

Sitting: This study was implemented in the Arabian Gulf University during the academic 2001-2002

Results: 80 students were examined on 11 stations for a total 110 minutes. . Two faculties were responsible in correcting the students' answer booklets. The mean station score given by 1st examiner was significantly higher than the mean station score given by 2nd examiner for most of the WOSCE stations. The mean total score of all stations for the 1st examiner was 97.13 ±SD 7.93 compared to 91.54 ± SD 10.00 for the 2nd examiner. This difference was statistically significant (P-Value = 0.00). Stations examined by the 1st examiner have coefficients of variation lower than that examined by the 2nd.

Conclusion: The WOSCE is an effective exam that is able to test certain clinical skills. The WOSCE exam had a medium to high inter-rater reliability. There is a mild to high consistency between the two raters.

Reliability of Watched Objective Structured Clinical Examination (WOSCE) Scores in Family Medicine Clerkship

Introduction

The College of Medicine and Medical Sciences (CMMS) at the Arabian Gulf University (AGU) is a community oriented medical school that adopts problem based learning (PBL) curriculum. Since its foundation in 1982, sixteen batches of doctors have been graduated. The school has 3 phases; phase 1, the premedical which extend for one year, phase 2 integrated organ system curriculum extending over three years. The main strategy of learning is problem based learning. Phase three is the clerkship phase. Students rotate for two years between different departments (Surgery, Medicine, Pediatrics, ObGyn, Psychiatry and Family Medicine. At the end of the clerkship phase and before the final MD exam, students spend 6 weeks of Family Medicine training in primary health care centers mainly at the students' own country of origin, since students come from different Gulf countries.

On completion of the rotation students are assessed by three different methods which include; supervisor's end of rotation as evaluation of students' activities during the clerkship, clinical encounter examination with real patients and the watched objective structured clinical examination (WOSCE). In the WOSCE a number of clinical scenarios (real and simulated patients) are video-filmed and presented to all the students at the same time. The students are required to respond to the various statements stated in the WOSCE booklet that is handed to them during the examination. The WOSCE is similar to the OSCE but with many advantages that it is less time consuming, needs less

manpower to conduct the exam and therefore less costly and less stressful to the students. Such advantages have been claimed for all objective clinical examinations [1].

Our earlier publication described in detail the WOSCE, its construction, implementation and explored its students' satisfaction [2].

To develop an ideal students' assessment method of clinical competencies is still controversial issue [3]. Reliability is a problem when using case examination in assessing clinical competencies of a medical student [1]. Students' performance could vary according to the type of examination method and the specificity of the medical problem encountered [4,5]. Also examiners differ in their judgments of students' clinical performances due to differences in focus and standards [3]. Therefore and in order to avoid validity and reliability problems students' assessment of clinical competencies needs more than one method of examination such as the OSCE and the WOSCE [6]. Many medical schools are using for clinical competency testing [3] different structured clinical examination format which has been claimed to be more valid and more reliable [3,7] However, Hamdy et al reported that traditional type of students' assessment by using long cases in clinical examinations is still being adopted by many medical schools [4].

The CMMS developed several students' assessment methods. This includes classical with simulated patients, WOSCE and direct observation clinical encounter examination (DOCEE). The later is performance-based examination aiming at evaluating students' clinical competences under direct observation during a clinical encounter with real patients [8]. These different examinations aimed at improving reliability and validity of students' clinical competencies [4].

Several studies showed that increasing the number of raters increases the reliability of the examination [1]. Therefore the authors recommend that for such examination probably one examiner is sufficient [1]. It was also reported that a common index for estimating the reliability of data collected in observational studies is the inter-observer agreement[9].

The aim of our study was:

1- To find out the examiners' degree of agreement on pass/fail decisions.

2- Ranking and overall examination reliability.

Methods

This study was implemented during the academic year 2001-2002. All year 6 students (N=80) who completed the family medicine clerkship rotation sat the WOSCE examination. This exam aimed at assessing students' communication skills, professional values, knowledge, diagnostic and management decisions. Also the exam tested the skills of providing of health education and ability to handle various forms such as growth chart and prescription writing. The examination consisted of 11 stations (table1).

For each station a scenario and a script was developed. It was video filmed using real patient (in 7 stations) and professional simulated patients (in 4 stations). The videos were observed by two family physicians using a checklist to test its content validity, clarity of sounds and picture, calculate the time and verify the quality of questions. The structured WOSCE booklet was also examined for its content and the answer sheets layout. The model answer booklet was designed after reviewing all model answers. The 80 students were seated in two halls supervised by one invigilator in each and support was given by Audio Visual technician to avoid any breakdown in the AV. system. Each hall was equipped with a full set of audiovisuals such as video camera, projector, loud speakers and a timer bell set to regularly ring at certain fixed time. To all the students the WOSCE booklet was distributed which contained information and questions about each station. The students were required to carefully watch and listen to the scenarios, read the instruction in the booklet and respond to the questions within the specified time. Each station extended for ten minutes to simulate the actual consultation time in a primary health care set up. The whole exam lasted for less than two hours. Two qualified family physicians that developed the exam were given the responsibility of marking

independently the answers and not to collude or alter the scores after discussion. Later the averages marks of both examiners for each station were considered to be the students' final mark. A score of 10 was given for each station and the total mark out of 100 was grouped into 5 categories according to the CMMS grade point average as follow; F= 0-59.9, D=60.0-69.9, C=70.0-79.9, B= 80.0-89.9, A=9-.0-100.

Statistical Analysis

Data were stored and analyzed using the Statistical Package for Social Sciences SPSS version 11.5. Relevant descriptive statistics of the scores of both examiners were obtained for each station. The Pearson product-moment correlation was used to determine the extent to which ranks given by both examiners in each station were similar. The inter-rater reliability of the paired examiners was estimated on each station, using the interclass correlation coefficient (ICC). To measure and analyze the degree of agreement between the two examiners we calculated the overall percent agreement and obtained the chance-corrected agreement (kappa statistics) for each station. These measures of agreement were obtained based on categorizing the student's scores as passing or failing and by categorizing the scores to 5 categories (F= 0-5.9, D=6.0-6.9, C=7.0-7.9, B= 8.0-8.9, A=9-.0-10). The paired samples t-test was used to test the equality of the mean scores of the two examiners in each station.

Results

The mean station score given by 1st examiner was significantly higher than the mean station score given by 2nd examiner for all stations except for station 7 (P value = 0.296), station 9 (P value = 0.052) and station 11 (P value = 0.496) the means were similar. The station standard deviation was similar for both examiners in all stations. The mean total score of all stations (out of 110) for the 1st examiner was 97.13 ±SD 7.93 compared to 91.54 ± SD 10.00 for the 2nd examiner. This difference was statistically significant (P-Value = 0.00). Stations examined by the 1st examiner have coefficients of variation lower than that examined by the 2nd examiner except for station 11(Table 2).

Table 3 presents the Pearson correlation coefficients, Interclass correlation coefficients, kappa values and overall percent agreement between the two examiners. Interclass correlation coefficients (inter-rater reliabilities) between 1st examiner and 2nd examiner ratings ranged from 0.35 to 0.94 with a mean correlation of 0.60. Also, Pearson correlations between examiners ranged from 0.38 to 0.94 with a mean of 0.65. The kappa values calculated based on categorizing to 5 categories ranged from 0.14 to 0.54 with a mean of 0.32, which are considered bad to good. But, overall percent agreement ranged from 0.31 to 0.96 with a mean percent agreement of 0.63, which are considered satisfactory to excellent.

The kappa values for pass/fail are higher than those calculated using 5 categories and ranged from 0.29 to 1.00 with a mean of 0.64, which are considered satisfactory to excellent. Also, overall percent agreement ranged from 0.85 to 1.00 with a mean of 0.96, which are considered excellent.

Discussion

Reliability is defined as the extent to which examinee scores are stable or reproducible across different but similar samples of item, raters, testing sites, time of the day, patients etc **[**10**].** Since inter-rater reliability between the two examiners was good we could consider that there was a high consistency between the examiners.

Stalenhoef-Halling et al, 1990, have reported that generalizability coefficients for using one versus two raters show only marginal differences. May be the subjectivity involved from raters' disagreement has a limited influence on the overall precision of the test [11].

For a written test it is better to have raters rating one question for all the examinees than the same rater rates all questions for a limited number of examinees [10].

The study found that the inter-rater reliability of the WOSCE exam was high and there was a consistency between the two examiners. A study from the University of Toronto showed that there is a consistency in the judgments of examiners in defining the cutting scores in an OSCE examination [12].

If the final results of the students in the WOSCE were considered to be pass or fail, as being implemented in other clinical examinations, then our study have shown that there is a high agreement between the two raters. However if the WOSCE's final grade are categorized into a letter grade system ranking (e.g. A, B, C, D, E ,F) then the inter-rater agreement is considered moderate to low. Hence it is recommended that the model answers for the station to be as much as possible constructed in structured format to reduce subjectivity.

Structured examination can achieve reliabilities if sufficient resources were provided [13].

This study showed that WOSCE which has the advantage of being less demanding concerning cost involved in manpower "i.e. examiners", measure objectively important core competencies relevant to the discipline of family medicine is also statistically reliable.

Table 1: Type of WOSCE Stations.

1-Case recognition, Down's syndrome and understanding its mode of hereditary transmission with genetic problem.
2-Childhood vaccination.
3-Diagnosing a case of chickenpox and identifying mode of transmission.
4-Diagnosing and managing diaper rash.
5-Plotting growth chart and recognizing failure to thrive and defining its causes.
6-Diagnosing and managing a diabetic patient.
7-Diagnosing and managing thyrotoxicosis.
8-Traveler's health education.
9-Diagnosing and managing psychiatric condition (depression).
10-Writing a prescription for a pathological condition using a prescription form.
11-Understanding causes of consultation failure due to failure of doctor-patient relationship.

Table 2: Description of the WOSCE stations and paired samples t-test P-Values

Station	Examiner	Mean (10)	SD*	CV† (%)	Lowest Score	Highest Score	P-Value
1	1st	9.70	0.71	7.32	6.00	10.00	0.028
	2nd	9.56	0.94	9.83	6.00	10.00	
2	1st	8.66	1.63	18.82	2.00	10.00	0.000
	2nd	7.33	1.97	26.88	2.00	10.00	
3	1st	8.35	3.12	37.37	0.00	10.00	0.000
	2nd	7.72	3.15	40.80	0.00	10.00	
4	1st	8.61	2.34	27.18	0.00	10.00	0.000
	2nd	7.49	2.13	28.44	0.00	10.00	
5	1st	8.24	1.24	15.05	4.00	10.00	0.000
	2nd	7.64	1.53	20.03	3.00	10.00	
6	1st	8.59	1.09	12.69	5.00	10.00	0.002
	2nd	8.13	1.24	15.25	5.00	10.00	
7	1st	9.79	1.19	12.16	0.00	10.00	0.296
	2nd	9.74	1.22	12.53	0.00	10.00	
8	1st	9.36	1.14	12.18	3.00	10.00	0.000
	2nd	8.47	1.74	20.54	2.00	10.00	
9	1st	9.98	0.22	2.20	8.00	10.00	0.052
	2nd	9.89	0.45	4.55	8.00	10.00	
10	1st	8.71	0.8	9.18	7.00	10.00	0.002
	2nd	8.34	1.03	12.35	4.00	10.00	
11	1st	7.15	2.82	39.44	0.00	10.00	0.496
	2nd	7.24	2.55	35.22	0.00	10.00	

* Standard deviation.

† Coefficient of variation.

Table 3: Inter-rater reliability of the paired examiners

Station	Pearson Correlation*	ICC‡	% Agreement	Kappa
1	0.77	0.74	0.90	0.54
2	0.63	0.49	0.36	0.15
3	0.90	0.88	0.71	0.47
4	0.78	0.69	0.48	0.23
5	0.47	0.42	0.49	0.30
6	0.41	0.38	0.31	†
7	0.94	0.94	0.94	†
8	0.54	0.42	0.54	0.14
9	0.48	0.37	0.96	0.39
10	0.38	0.35	0.51	†
11	0.90	0.90	0.74	†

* All statistically significant at 0.01 level (two tailed).

‡ Interclass correlation coefficient: two-way random effects, parallel model, single measure.

†Kappa statistics cannot be computed

References

1. Wass V, Jolly B. Does observation add to the validity of the long case? Medical Education 2001; 35:729-34.
2. Alnasir FA. The Watched Structure Clinical Examination (WASCE) as a tool of assessment. Saudi Med J. 2004 Jan; 25(1):71-4.
3. Meadow R. The structured exam has taken over. BMJ 1998; 317:704-5.
4. Hamdy H, Parsad K, Williams R, Salih F. Reliability and validity of the direct observation clinical encounter examination (DOCEE). Medical Education 2003; 37:205-12.
5. Ferrell BG. Demonstrating the efficacy of the patient logbook as a program evaluation tool. Acad Med 1991; 66:49-51.
6. Cohen R, Rothman AI, Poldere P, Ross j. Validity and generalizability of global ratings in an objective structured clinical examination. Acad Med 1991; 66:545-8.
7. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Medical Education 1979; 13:41-54.
8. Abouna GM, Hamdy H. The integrated direct observation clinical encounter examination (DOCEE). An objective assessment of students' clinical competence in a problem based learning curriculum. Med Teach 1999; 21:67-72.
9. Mitchell SK. Interobserver agreement, reliability and generalizability of data collected in observational studies. Psychol Bull 1979; 86:376-90.
10. Van Der Vleuten CPM, Norman GR, DE Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. Medical Education 1991; 25:110-18.
11. Stalenhoef-Halling BF, Jaspers TAM, Fiolet JFBM, Van der Vleuten CPM. (1990). The feasibility, acceptability and reliability f open-ended questions in problem-based learning curriculum In: Teaching and assessing clinical competence. (ed.by W. Bender, RJ Hiemstra, AJJA Scherpbier & RP Zwierstra). Boekwerk Publ., Groningen.

12. Schwiebert P, Davis A. Increasing inter-rater agreement on a family medicine clerkship oral examination—a pilot study. Fam Med 1993; 25:182-5.

13. Wass V, Wakeford R, Neighbour R, Van der Vleuten C. Achieving acceptable reliability in oral examinations: an analysis of the Royal College of General Practitioners membership examination's oral component. Med Educ 2003; 37:126-31.