Journal of General Practice
ISSN: 2329-9126

Schoenmakers and Ryssaert, J Gen Pract 2013, 2:1
DOI: 10.4172/2329-9126.1000134

Research Article Open Access

Using a 'Yellow Card' in the Objective Structured Clinical Exam: Does it Add to the Identification of Problem Postgraduate Trainees in General Practice? An Exploratory Study to Identify High Risk Trainees

Birgitte Schoenmakers[1,2]* and Lynn Ryssaert[3]

[1]Department of Public Health and Primary Care, Academic Centre of General Practice, University Leuven, Leuven, Belgium
[2]Academic Teaching Practice, Department of Public Health and Primary Care, Belgium
[3]Academic Centre of General Practice, Leuven, Belgium

Abstract

Background: The Objective Structured Clinical Examination was designed 30 years ago by Harden ea. It is used to assess various components of medical competence. The OSCE is considered as a reliable and objective tool to evaluate clinical competences in standardized patient encounters. Although, reliability, validity and reproducibility an OSCE remain subject of debate. These days the question arises if a compensatory or an additional rating is advisable for the final pass-fail decision.

Aim: The aim of the study is to add to the identification of high risk postgraduate trainees in general practice by means of 'a yellow card system' (red flagging).

Method: During 8 OSCE-sessions, including 354 GP-trainees, observers were asked to deal a yellow card in case of 'alarming performance'. These acts were defined as dramatic or dangerous shortcomings on three levels: theoretical, practical and behavioral level.

Result: During three academic years, involving 354 trainees, only 41 yellow cards were dealt. One single observer was responsible for one quarter of all allocations. During two sessions half of all cards were dealt. Trainees remembered with a yellow card were more likely to underperform on all assessments except on the internship. During their internships, trainees with a yellow card did not show remarkable or alarming behavior.

Conclusion: Flagging alarming events during the OSCE does not identify high risk trainees. The idea of 'flagging' is not be abandoned but moved to other assessment situations.

Keywords: Objective structured clinical examination; Medical education; Postgraduate trainees; Assessment; General practice

Introduction

The Objective Structured Clinical Examination (OSCE) was designed 30 years ago by Harden and Gleeson [1]. It has been used in many countries all over the world to assess various components of medical competence [2,3]. The OSCE is considered as a reliable and objective tool to evaluate clinical competences in standardized patient encounters [4]. Although this test is generally accepted as a high stake assessment process, reliability, validity and reproducibility of an OSCE remains subject of discussions and concerns [5-7]. On trainee level, there is an ongoing debate on the efficacy of evaluating skills using an item-checklist [5]. Some trainees perform in a rather unstructured or even chaotic way and start guessing to hit the right 'keywords' in answer to the particular item. These students will, unintentionally and undeserved obtain a high score [8]. Second, the competence level of the examinees might influence the score in a disproportional way: poor performing trainees might score higher than more qualified trainees because the latter sometimes skip steps in the explicitation of the reasoning process [9,10].

On observer level, there are concerns about the variability of the observations. Observers are preferably experienced (practice) teachers but can also be trained patients [8,11]. Therefore a training, with firm instructions, prior to an OSCE observation is recommended to reduce rater inconsistency [9]. Finally, the performance of trainees on an OSCE is also limited by the artificial circumstances which might negatively affect both trainee and patient [12].

To obtain a reliable score, multiple clinical encounters are required [8]. While rotating in an OSCE, trainees are assessed on different skills like history taking, communication skills and clinical examination [9,11]. An OSCE is mostly composed of 6 to 20 stations [4]. Besides the varying amount of OSCE stations, the ideal length of each encounter is debatable [8]. Finally, most OSCE's take about 2 hours of rotation [4]. The OSCE aims at an objective evaluation of each participant. Therefore item checklists are developed as formal tools to score the performance of the individual trainee in each station. The final score is then calculated from the score on the item checklist, a separate appraisal on general vocational skills and a global score given by the observer [7].

In spite of all intrinsic and extrinsic 'safety checks', a discussion remains on the accountability of an OSCE in estimating trainee's performance in patient encounters. There are sufficient cases reporting on contradictory results between OSCE- and residency or internship-performance [13]. Therefore the question imposes whether a

***Corresponding author:** Birgitte Schoenmakers, Department of Public Health and Primary Care, Academic Centre of General Practice, University Leuven, Leuven, Belgium, Tel: +32 16 37 72 90; E-mail: Birgitte.schoenmakers@med.kuleuven.be

Received July 16, 2013; **Accepted** November 05, 2013; **Published** November 10, 2013

Citation: Schoenmakers B, Ryssaert L (2013) Using a 'Yellow Card' in the Objective Structured Clinical Exam: Does it Add to the Identification of Problem Postgraduate Trainees in General Practice? An Exploratory Study to Identify High Risk Trainees. J Gen Pract 2: 134. doi: 10.4172/2329-9126.1000134

compensatory or an additional rating, addressing behavior and overall impression of competence, is advisable for the final pass-fail decision [13].

This study explores the added value of the introduction of a yellow card system. This intervention has no precedents in medical education assessment and has been developed by analogy with the flagging system in sports competition. In concrete, observers of an OSCE were asked to deal a yellow card under certain conditions. The hypothesis is that a yellow card system could help to identify the high risk (underperforming) trainees. The observer was asked to deal a symbolic yellow card to a trainee who presented with alarming acts during the patient encounter. The hypothesis is tested by linking the dealing of a yellow card to the assessment of the internship and the score on the other master exams.

Methods

Study population

After their graduate medical education trajectory, trainees opt for further specialization. The trainees in this study are postgraduate trainees in general practice. They fulfill a two year internship in a peripheral family practice. The final competence assessment in their graduation year includes an oral examination, a knowledge test, an OSCE and the evaluation of their internship.

This study took place during the academic years 2008-2009, 2009-2010 and 2010-2011. In total 354 individual trainees (distributed over these three academic years) took part in the OSCE and were included in the study. Since the study took part in the formal exam schedule, all trainees participated and there was no dropout.

Composition of the OSCE

The OSCE in this study case consisted of 20 stations. Each clinical encounter took 8 minutes. The yellow card system was tested during 8 sessions of the OCSE. These 8 sessions were distributed over three academic years. The composition of the OSCE varied per session to avoid the dissemination of the content by the trainees.

The observers were all experienced clinicians (more than 5 years after final graduation) and academic teachers (also clinicians). Before participation to an OSCE, observers received a training given by the assessment staff. Observers were asked to rate the performance by means of the traditional item checklist for each station. Besides, observers also provided a global score on each performance.

The OSCE was composed (for more than 15 years) by an experienced assessment staff. Each station was designed following a formal script and protocol and validated by the staff members. Subsequently, scenarios were rejected, accepted or revised. Inter-observer variance was tested by double observations during the exam. Psychometric testing was performed after the exam to guarantee validity and reliability of the OSCE.

Study design and outcome

Observers were asked to deal a yellow card in case trainees presented with 'alarming acts'. These acts were defined as dramatic or dangerous shortcomings addressing three levels: theoretical, practical and behavioral level. The affected trainee was not informed about this action to avoid interference with the ongoing formal examination. The dealing of a yellow card was an action for research purposes only without any impact on the final OSCE-score. This intervention was considered as subject of an exploratory study.

In concrete, yellow cards were reserved for four different cases

1. when the trainee performed with dangerous or even life threatening clinical interventions.

2. when the trainee performed with unethical behavior including incorrect advice or proposals.

3. when a trainee behaved in a rude, uninterested or impolite way.

4. when 'general shortcomings' were observed

Besides this structured and standardized classification, observers were also asked to document and motivate their 'yellow card' decision in a free text field.

The outcome measures were subsequently defined based upon the four aforementioned conditions. The features of the study-population were included as co-variables: finals score on the OSCE and score on the particular station, scores on other exams and performance during internship.

Analyses

The analyses were performed with SPS19 (Statistical Package for the Social Sciences, version 19, IBM, US). Univariate and bivariate analyses were performed to describe the features of the study population and the primary outcome measures. Subsequently, a correlation analysis refined the understanding of the relation between the final score on the OSCE, on other exams and on the internship and the occurrence of a yellow card. Further multivariate analysis was found to be unreliable due to the low occurrence of a yellow card as compared to the size of the study population.

Ethical approval

This study was conducted without impact on trainees' functioning or performing. When trainees enroll at the university they agree by intrinsic contract that anonymous data concerning their learning and assessment activities can be used for research purpose. Second, the policy of the Medical Ethical Advisory Board of the faculties involved follows the national legislation in research matters. A request for research approval is therefore not required when no authentic or true patient-doctor contact is involved.

Results

Over a period over the three academic years with in total 8 OSCE-sessions, a total of 41 yellow cards was dealt by 19 individual observers. Table 1 show the frequency of appearance of the stations where at least once a yellow card is dealt.

Station	Frequency (n)	Station	Frequency (n)
Anticoagulant prescription	8	Adolescent problem	6
Ancle pain	8	Abdomen pain	6
Headache	8	Cardio pulmonary resuscitation	6
Low back pain	8	Menopause diagnosis	6
Oral anti-conception prescription	8	Oncology referral	6
anti-conception communication	8	Thyroid diagnosis	6
Dementia diagnostic stage	2	Palliative collaboration	3

Table 1: Distribution of the stations were yellow cards were dealt (=n).

One single observer was responsible for almost a quarter of the interventions. Four trainees received more than one yellow card during the same OSCE: two trainees received two cards; one trainee received three cards and one trainee up to four cards.

The situations leading to a yellow card were described as 'inappropriate policy strategy, inadequate communication, lack of global consultation skills, shortcomings in diagnostic skills, improper behavior, insufficient knowledge or poor competence in clinical examination'. The two latter were by far the main reasons for dealing a yellow card.

In the stations 'low back pain', 'ankle pain', 'Cardio pulmonary resuscitation', 'anticoagulant therapy' and 'oncology' (adjusted for their relative share in all sessions) a yellow card was dealt in respectively 27%, 15%, 9%; 9% and 9% of all trainee-patient encounters. Strikingly, two OSCE- sessions yielded together 50% of all yellow cards distributed.

Following the hypothesis that the 'yellow card' intervention contributes to the identification of poor performing trainees, correlations with the other assessments were explored (Table 2). Scores on the oral exam, the master thesis and the knowledge test were taken into analyses and compared to the OSCE-score and the allocation of a yellow card.

Initially, 33 trainees received one or more yellow cards. One trainee was excluded for final analyses since he repeated graduation year with extra-ordinary results on all assessments. He failed and dropped out the year before due to due to personal problems (Table 2).

The final assessment scores of each individual trainee were rather closely related on intra-trainee level. But, bivariate analysis showed a significant difference between the overall assessment scores of both groups (with and without yellow card) except for the internship. Scores of the internship assessment were comparable in both groups. Trainees who received a yellow card were more likely to have a lower overall assessment scores as compared to the trainees without yellow card.

A correlation analysis confirmed the inverse relationship between the allocation of a yellow card and overall assessment scores; except for the internship (lower scores are related to the allocation of a yellow card).

A manual tracking of the internship assessment files of the 'yellow card trainees' did not yield any remarkable or alarming comment on any level made by the supervisors.

	mean				
Type of exam	**total group**	**yellow card**	**no yellow card**	**Spearman**	**p-value**
	n	*n*	*n*	r	
master thesis	14,60	13,81	14,68	0,159	0,003
	346	*32*	*314*		
oral exam	13,28	12,34	13,37	0,145	0,007
	352	*32*	*320*		
Knowledge	14,99	14,28	15,06	0,109	0,041
	353	*32*	*321*		
clinical & communicational skills-OSCE	13,77	12,16	13,93	0,218	0,001
	352	*32*	*320*		
internship	15,61	15,13	15,66	0,101	0,059
	351	*32*	*319*		

Table 2: Mean scores on other exams in graduation year, the correlation between allocation of a yellow card and other test scores.

Discussion

In this exploratory study the contribution of a 'yellow card' intervention to the identification of high risk postgraduate trainees in general practice was tested. During three consecutive academic years, including 354 trainees and eight OSCE-sessions, only 41 yellow cards were dealt. Strikingly, one single observer was responsible for one quarter of the allocations. During two sessions half of all yellow cards were dealt to poor performing trainees. Trainees remembered with a yellow card were more likely to underscore on all assessments except on the internship. During their internships, trainees with a yellow card did not show remarkable or alarming behavior.

In this study, the allocation of a yellow card seemed not related to the reporting of a similar performance or behavior during internship. An OSCE is expected to assess clinical and professional competencies. The content and format of an OSCE represent an objective and structured simulation of reality. Therefore the OSCE is complementary to the internship assessment [14,15]. Several considerations need to be addressed. First, the validity of the OSCE in assessing clinical competencies needs to be questioned. Debate on the added value of an OSCE as compared to oral exams (from jury exams to the so called long case) remains active [7,16,17]. Opponents of the OSCE dispute the content reliability and veracity. Second, evaluation of individual internships is influenced by emotional involvement of supervisor and trainee [18]. A subjective and unintentional higher appreciation of the trainees' performance is therefore inevitable. Besides, a poor performing trainee could be the result of a poor training under the responsibility of the concerned supervisor. For that reason, supervisors are reluctant to give a low(er) score on internship assessment. Third, factors inherent to the nature of the OSCE could influence the observers' appreciation of the trainees' performance [17]. The characteristics of both the simulated patient and the observer (gender, psychosocial and cultural features) and the rotation order of the OSCE are known to interfere with trainees' performance and thus with the final test score [19,20].

To address the above considerations, the OSCEs of the past 10 years organized by our universities were revisited. Validity, reliability and reproducibility of these examinations were found to be good to very good. Therefore, it is acceptable to put that the OSCEs included in this study have similar test qualities as the previous sessions. Considerations related to the particular OSCE context can be rejected. Above, the proportion of failed versus successful trainees remained stable over the past 13 years.

Considering the above argument, a particular observer effect or interaction cannot be ruled out. Indeed, as noted in this study one observer accounted for a quarter of all yellow card interventions. Second, most cards were dealt in two sessions while other sessions were 'yellow card-free'. Perhaps some observers believed they were expected to document or motivate their scoring by means of a yellow card. Either, observers weren't well instructed and felt unsure or less skilled in evaluating trainees' performance. Another plausible explanation could be that the opportunity (yellow card) created the action (dealing it). Remarkably, the allocation of a yellow card did not necessarily led to a fail or a poor score on the particular encounter. Some authors therefore advocate the involvement of the observers throughout the developing of an OSCE [21].

Finally, it was observed that 'yellow card' trainees scored less on all assessments as compared to trainees without yellow card. But on the other hand, these trainees were not appointed with inappropriate or

alarming behavior during their internships. As part of the assessment and training all trainees conducted several supervised consultations with immediate and postponed feedback (as part of the educational program). It is plausible that internship scores are not always reliable although qualitative assessment and feedback are known to be trustworthy [18]. A manual tracking of all assessment files yielded three to five trainees a year noticed with alarming behavior or problematic performance. None of these trainees was remembered with a yellow card during the OSCE.

Strengths and limitations

The strength of the study certainly lies in the implementation of the intervention in the regular assessment process. All trainees and all sessions of the past three years were included. The study was interrupted after the third trial year because of the lack of an intervention effect.

Strength of the study is that the often discussed principle of 'red flagging' was addressed. Flagging in case of critical issues and alarming situations is believed to be an efficient strategy in the identification of trainees at risk [22,23]. Occasionally, OSCE observers indicate that they ticked all items (corresponding to a high test score) but that they also observed an alarming shortcoming with the trainee. This study demonstrated that flagging by means of a yellow card did no add value to the final pass/fail decision.

The study also has some considerable limitations. Due to the small amount of yellow cards dealt, profound or multivariate analysis was impossible. But, since no correlation between a yellow card and functioning during internship was found, further exploration seems needless. Another weakness of this study is the lack of a follow up of trainees provided with a yellow card. Indeed, trainees were not informed about the intervention and the yellow card didn't affect their final test result since the study was set up in an exploratory design. Ideally, these trainees were tracked and offered feedback [22].

Conclusions and implications for research and practice

Counting on the test quality of an OSCE, the allocation of a yellow card seemed rather dependent on the observer and the 'flagging' opportunity. Therefore, flagging alarming behaviors or problematic performance during simulated patient encounters does not identify high risk postgraduate trainees in general practice. Since one observer was found to be responsible for half of the cards dealt, we particularly learnt that the objective, structured and standardized character of each assessment merits a permanent attention. Repeated training of and feedback to observers and simulated patients are essential to maintain a high assessment quality. But, considering that 'yellow card students' were more likely to underscore on other assessments, the idea of 'flagging' is not abandoned. Further research should therefore focus on the simultaneous use of this concept in the different assessments.

References

1. Harden RM, Gleeson FA (1979) Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 13: 41-54.
2. Sibert L, Mairesse JP, Aulanier S, Olombel P, Becret F, et al. (2001) Introducing the objective structured clinical examination to a general practice residency programme: results of a French pilot study. Med Teach 23: 383-388.
3. Khan KZ, Ramachandran S, Gaunt K, Pushkar P (2013) The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. Med Teach 35: e1437-1446.
4. Harden V, Harden RM (2003) OSCE Annotated Bibliography with Contents Analysis. BEME Guide no 17.
5. Turner JL, Dankoski ME (2008) Objective structured clinical exams: a critical review. Fam Med 40: 574-578.
6. Lin CW, Clinciu DL, Swartz MH, Wu CC, Lien GS, et al. (2013) An integrative OSCE methodology for enhancing the traditional OSCE program at Taipei medical university ospital - a feasibility study. BMC Med Educ 13: 102.
7. Pell G, Fuller R, Homer M, Roberts T; International Association for Medical Education (2010) How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. Med Teach 32: 802-811.
8. Whelan G, Boulet J, McKinley D, Norcini J, van Zanten M, et al. (2005) Scoring standardized patient examinations: lessons learned from the development and administration of the ICFMG Clinical Skills Assessment (CSA). Med Teach 27: 200-206.
9. Boursicot K, Roberts T (2005) How to set up an OSCE. The clinical teacher 2: 16-20.
10. Hodges B, McNaughton N, Regehr G, Tiberius R, Hanson M (2002) The challenge of creating new OSCE measures to capture the characteristics of expertise. Med Educ 36: 742-748.
11. Van Nuland M, Hannes K, Aertgeerts B, Goedhuys J (2005) Educational interventions for improving general practice trainees communication skills in the clinical consultation. Cochrane Database of Systematic Reviews 4: 1-12.
12. Bergus GR, Woodhead JC, Kreiter CD (2009) Trained lay observers can reliably assess medical students' communication skills. Med Educ 43: 688-694.
13. Chesser AM, Laing MR, Miedzybrodzka ZH, Brittenden J, Heys SD (2004) Factor analysis can be a useful standard setting tool in a high stakes OSCE assessment. Med Educ 38: 825-831.
14. Shumway JM, Harden RM; Association for Medical Education in Europe (2003) AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. Med Teach 25: 569-584.
15. Hendrickx K, De Winter B, Tjalma W, Avonts D, Peeraer G, et al. (2009) Learning intimate examinations with simulated patients: the evaluation of medical students' performance. Med Teach 31: e139-147.
16. Ponnamperuma GG, Karunathilake IM, McAleer S, Davis MH (2009) The long case and its modifications: a literature review. Med Educ 43: 936-941.
17. Hodges B (2003) Validity and the OSCE. Med Teach 25: 250-254.
18. Murphy DJ, Bruce DA, Mercer SW, Eva KW (2009) The reliability of workplace-based assessment in postgraduate medical education and training: a national evaluation in general practice in the United Kingdom. Adv Health Sci Educ Theory Pract 14: 219-232.
19. Humphrey-Murto S, Touchie C, Wood TJ, Smee S (2009) Does the gender of the standardised patient influence candidate performance in an objective structured clinical examination? Med Educ 43: 521-525.
20. Blaskiewicz RJ, Park RS, Chibnall JT, Powell JK (2004) The influence of testing context and clinical rotation order on students' OSCE performance. Acad Med 79: 597-601.
21. Wilkinson TJ, Frampton CM, Thompson-Fawcett M, Egan T (2003) Objectivity in objective structured clinical examinations: checklists are no substitute for examiner commitment. Acad Med 78: 219-223.
22. Pell G, Fuller R, Homer M, Roberts T (2012) Is short-term remediation after OSCE failure sustained? A retrospective analysis of the longitudinal attainment of underperforming students in OSCE assessments. Med Teach 34: 146-150.
23. van Mook WN, van Luijk SJ, O'Sullivan H, Wass V, Schuwirth LW, et al. (2009) General considerations regarding assessment of professional behaviour. Eur J Intern Med 20: e90-e95.