Can student tutors act as examiners in an objective structured clinical examination?

Jean-François Chenot,1 Anne Simmenroth-Nayda,1 Alexandra Koch,1 Thomas Fischer,1 Martin Scherer,1 Birgit Emmert,2 Beate Stanske,3 Michael M Kochen1 & Wolfgang Himmel1

CONTEXT The dissemination of objective structured clinical examinations (OSCEs) is hampered by requirements for high levels of staffing and a significantly higher workload compared with multiple-choice examinations. Senior medical students may be able to support faculty staff to assess their peers. The aim of this study is to assess the reliability of student tutors as OSCE examiners and their acceptance by their peers.

METHODS Using a checklist and a global rating, teaching doctors (TDs) and student tutors (STs) simultaneously assessed students in basic clinical skills at 4 OSCE stations. The inter-rater agreement between TDs and STs was calculated by kappa values and paired t-tests. Students then completed a questionnaire to assess their acceptance of student peer examiners.

RESULTS All 214 Year 3 students at the University of Göttingen Medical School were evaluated in spring 2005. Student tutors gave slightly better average grades than TDs (differences of 0.02–0.20 on a 5-point Likert scale). Inter-rater agreement at the stations ranged from 0.41 to 0.64 for checklist assessment and global ratings; overall inter-rater agreement on the final grade was 0.66. Most students felt that assessment by STs would result in the same grades as assessment by TDs (64%) and that it would be similarly objective (69%). Nearly all students (95%) felt confident that they could evaluate their peers themselves in an OSCE.

CONCLUSIONS On the basis of our results, STs can act as examiners in summative OSCEs to assess basic medical skills. The slightly better grades observed are of no practical concern. Students accepted assessment performed by STs.

KEYWORDS educational measurement/methods; *education, medical, undergraduate; *peer review; *teaching; Germany; *students, medical; observer variation.

Medical Education 2007: 41: 1032–1038
doi:10.1111/j.1365-2923.2007.02895.x

INTRODUCTION

The objective structured clinical examination (OSCE) is considered to be one of the most adequate formats for assessing clinical skills and has gained widespread distribution since its introduction.1 Considerable variability has been observed in how medical schools assess clinical skills.2 Although OSCEs are used as substantial part of the licensure process in some countries, the OSCE has never become very popular in other countries outside small reform universities or ephemeral attempts at some larger medical schools. High staff requirements, high time expenditure and considerable costs have prevented its introduction.3,4 Multiple-choice and oral examinations are still the most commonly used assessment formats.

The new curriculum for medical schools introduced in Germany in 2004 places more emphasis on clinical skills.5 As a result, we introduced a clinical skills course, which combines training in procedural skills
and history taking. We used student tutors (STs) to ensure small group sizes and optimal instruction for the practical exercises. We also assessed student performance with a summative OSCE. Because our capability to perform an OSCE for approximately 200 students twice a year is constrained by limited human resources, we looked for additional support in student performance assessment. Expert examiners (i.e. medical teachers) might not be necessary to assess students and medical students who were involved in the teaching seemed to be the perfect choice for the task. Therefore, we decided to evaluate the reliability and acceptability of STs for assessment of students in an OSCE.

Methods

Participants

Having completed an interdisciplinary course entitled ‘Introduction to Clinical Skills’, Year 3 medical students at the University of Göttingen were assessed with a summative OSCE (10 stations) in spring 2005. The OSCE is only 1 of the various components contributing to the final grade; therefore students cannot fail the examination.

A teaching doctor (TD) and an ST rated all the students at 4 stations. The stations had been used and refined previously, with good inter-rater correlations ranging from 0.7 to 0.8. For the standardised rater training, we used sample videos and gave detailed instructions on how to assess students’ performance. We trained 20 medical students in Years 4 and 5 who had assisted as STs during the course and 25 TDs to rate 1 or 2 stations each. Student tutors were contracted to the faculty of medicine and received €10.52 per hour for teaching and rating. Becoming an ST is popular and is considered to be an excellent opportunity to refine one’s own clinical skills.

Teaching doctors consisted of faculty staff and general practitioners associated with the faculty. They were not remunerated. Senior medical students and members of a local lay theatre group were recruited and trained as standardised patients.

Assessment

Objective structured clinical examinations frequently use the categories ‘pass’ or ‘fail’ as assessment criteria. Federal regulation of medical examination in Germany, however, requires a grade on an ordinal scale ranging from 1 to 5. On this scale, excellent (1), good (2), fair (3), sufficient (4) and fail (5) correspond to fixed percentages which have been set by the faculty. For example, a ‘1’ represents > 90% of the total score and a ‘5’ equates to < 50% of the total score. In our study, student performance was assessed with a checklist, which consisted of either single-choice items (e.g. student asked about smoking: yes or no) or Likert scales (e.g. student–patient interaction: good or fair or poor). These checklist items were added together to calculate a total score for each individual station. This score was then translated into a grade from 1 to 5 to generate the checklist rating (CR). In addition, each student received a global rating (GR) at each station, which was given on the same ordinal scale of 1–5. The GR represents an overall impression of the student’s performance and is needed to calibrate the CR for aspects that are not properly taken into account by the checklist. The final grade consists of a combination of the equally weighted CR and GR.

Students received a questionnaire after completing the OSCE to evaluate both the OSCE itself and to
estimate whether being evaluated by an ST would have made any difference. They were aware that the assessment was being performed by both a TD and an ST and that only the TD’s assessment was taken into account for the grade awarded.

**Stations on the OSCE**

Table 1 briefly describes the OSCE stations. Each task had to be accomplished in 5 minutes.

**Statistical analysis**

The average grade as a mean value (and standard deviation) was calculated for each station. Teaching doctor ratings were subtracted from ST ratings to obtain the average difference. Negative values indicate that STs had a tendency to give better grades than TDs and vice versa. Paired $t$-tests were used to establish whether there was a significant difference between ST and TD ratings. In a few instances, single items were missing. To ensure complete rating, we always assumed the best possible rating for the missing item on the checklist in favour of the students.

To compare inter-rater agreement on single and composite items, we first showed raw agreement and then calculated kappa values.\(^\text{10}\) If more than 2 categories are used, weighted kappa is appropriate to adjust for the degree of discrepancy. For interpretation of kappa values, we used the classification proposed by Landis and Koch (0.01–0.2 = slight agreement; 0.21–0.4 = fair agreement; 0.41–0.6 = moderate agreement; 0.61–0.8 = good agreement; 0.81–1.0 = very good agreement).\(^\text{11}\) We also checked for inter-rater agreement on each item of the checklist (data not shown). Data were processed using SAS 8.2.

**RESULTS**

All 214 Year 3 students participated in the OSCE. Their average age was 24.3 years (standard deviation [SD] ± 1.2) and 111 (52%) of them were male. Assessments made by TDs and STs were available for at least 98% of the students. On average, each TD assessed 34.2 (SD ± 9.6) students and each ST 42.7 (SD ± 9.3) students.

**Assessment of student performance and inter-rater agreement**

The results of the CRs, GRs and final grades are shown in Table 2. The average grades awarded by TDs and STs were very close to each other and most of the differences were not statistically significant. Student tutors had a tendency to give slightly better grades at some stations, which was most pronounced at the electrocardiogram (ECG) station. The GR was, on average, a little better from both STs and TDs except at the ECG station. The raw agreement between the 2 groups of raters ranged from 55% to 72%. Kappa statistics showed moderate to good inter-rater agreement ranging from 0.41 to 0.64. The best inter-rater agreement was obtained with the summative grade calculated from all 4 stations. Only 2 students failed the cardiovascular risk assessment station, 4 the ECG station, 7 the depression screening station and 4 the occupational assessment station. At the first 2 stations, there was perfect agreement.

| **Table 1 Description of the 4 objective structured clinical examination (OSCE) stations** |
| OSCE station | Description |
| Cardiovascular risk assessment | Instructions to students: a woman who had a high blood pressure reading in a pharmacy presents for the first time for assessment. Obtain a focused cardiovascular and a short social history. History taking is terminated with the announcement that you will take the patient’s blood pressure. Assessment criteria: language, structure, patient contact, items associated with cardiovascular risk (e.g. smoking history) |
| Electrocardiogram (ECG) | Instructions to students: a young man presents with chest pain. Obtain an ECG; explain what you are doing; give a short interpretation of the ECG; tell the patient the result (normal) |
| Depression screening | Instructions to students: a student presents with abdominal pain of 3 weeks duration. He has just had a normal gastroscopy. Obtain a focused psychosocial history. |
| Occupational assessment | Instructions to students: a young woman presents with wrist pain and requests a sick leave certificate. Obtain an occupational history; make a decision about whether or not you will sign a certificate; explain your decision to the patient. |

© Blackwell Publishing Ltd 2007. MEDICAL EDUCATION 2007; 41: 1032–1038
between TDs and STs about the students who failed the station; at the last 2 stations, the parties differed in only 1 or 2 instances.

The students’ perspective

More than 90% (195/214) of the students filled in the questionnaire. Nearly all of them (95%) had not participated in an OSCE before. About 93% agreed that the OSCE is an adequate assessment format to evaluate clinical skills. More than a third (40%, 78/195) favoured the OSCE, 34% preferred a multiple-choice type examination and 26% preferred other types of assessment, such as an oral examination. The majority (64%) felt that they would obtain the same grades, whether awarded by an ST or a TD, whereas 27% believed they would get a better and 8% a worse grade. That STs would be as objective as TDs was the belief of nearly 85% (179/195) and only a small proportion attributed TDs a higher (17%) or lower (14%) level of objectivity. The majority (165/195) agreed that STs had assessed them adequately and 185/195 felt confident that they would be able to assess younger students themselves in an OSCE (Table 3).

DISCUSSION

Summary of the main findings

Senior medical students can reliably rate students in an OSCE, with moderate to good agreement with TDs. There was essentially no difference in the assessment of basic medical skills by STs compared with TDs. The final grade calculated from the result of 4 OSCE stations had good inter-rater agreement. The majority of students believed that STs had assessed their performance in the OSCE adequately and as objectively as TDs.

Strengths and limitations

To our knowledge, this is the first published study on the quality and acceptability of students as raters on an OSCE to assess clinical skills in a large sample of medical students. However, peer assessment has been evaluated in other contexts. For example, Dannefer et al.12 and Lurie et al.13 studied peer assessment of classmates in the same year in medical schools. Only 1 study on dental students as peer examiners in an OSCE setting has been published.14 In this study, dental students and faculty members evaluated the examination of the mouth with a checklist. No significant differences were found.

Teaching doctors and STs were asked not to talk to each other about their evaluation. Nevertheless, there may still have been some consultation. As we cannot exclude the occurrence of such cross-contamination, it remains possible that the observed agreement between TDs and STs might be an overestimate.
Our findings cannot be generalised to all OSCEs as inter-rater reliability needs to be established for each OSCE station independently. That said, we have shown that senior students who are involved in the teaching of the skills can reliably assess their peers at OSCE stations with clearly operationalised items which do not require a high level of knowledge and understanding, after they have received appropriate training as raters. This might not be true for students who have not been previously involved in teaching.

Meaning of the results

In a review of peer assessment by medical students, a positive correlation between evaluations by residents and faculty was observed. In these studies, students usually evaluated their contemporaries in clinical settings. Although the students in our study evaluated less advanced medical students in OSCE settings, their ratings compared well with those of TDs. One reason for the acceptable correlation between student and faculty evaluations may be that our OSCE setting provided a clear structure, so that even less experienced raters had no, or only slight, difficulties in objectively evaluating their peers’ performance. Therefore, using senior students to support professional teachers also seems useful in the case of OSCEs and may overcome some barriers which have prevented their broader implementation up until now.

The best agreement was achieved at the cardiovascular risk assessment and the occupational assessment stations. At those 2 stations most items on the checklist had been dichotomised. Most items on the checklist for the depression screening station used Likert scales, which leaves more leeway for interpretation.

On average, student raters gave slightly better grades at the ECG station, which also yielded the lowest inter-rater agreement. The differences in global ratings probably mainly arose from the speed and self-assuredness of the students. Teaching doctors evaluated a little more rigorously, probably expecting more from the students than the basics that were taught in the course. At the ECG station, it was particularly hard to categorise student performance beyond ‘pass’ or ‘fail’ as most students were able to perform and interpret the ECG almost correctly. The checklist required raters to choose between correct and incorrect, which led to difficulties in how to deal with minor mistakes. Therefore, we will add a category ‘almost correct’ in future administrations.

Rudy et al. observed more lenient evaluation by peers compared with medical faculty. In some instances, we also observed ST ratings in favour of the students, but we also saw some in their disfavour. In any case, the differences are of no practical relevance. We have no evidence that STs awarded systematically better grades. Most students in our sample assumed that evaluation by STs would not affect their grade.

One might wish for higher inter-rater agreement than we observed. As inter-rater reliability depends on multiple factors, such as instruction of the raters, compilation of the checklist and involvement of the raters in the process of developing the station, one may look for improvements in each of these factors. However, optimising inter-rater agreement has to be balanced against the associated efforts and cost.

Some possible disadvantages of peer assessment in OSCEs have to be considered. Students may not have the same level of understanding as the teacher. They may also need clearer guidance in terms of what they should look for, may show bias towards their peers and may be reluctant to award low marks for poor work by fellow students. According to two recent studies, students differ in their ability to discriminate

Table 3 Student assumptions about their evaluation by tutor students and confidence about rating other students

<table>
<thead>
<tr>
<th>Item</th>
<th>Answers (%)</th>
<th>Mostly agree</th>
<th>Mostly disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I believe I have been adequately evaluated by student tutors</td>
<td>44</td>
<td>40</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>I am confident that I could rate younger students in an OSCE</td>
<td>43</td>
<td>32</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>I believe student tutors, compared with faculty, evaluate…</td>
<td>17</td>
<td>69</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>I believe my assessment by student tutors, compared with faculty, is…</td>
<td>27</td>
<td>64</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

OSCE = objective structured clinical examination

© Blackwell Publishing Ltd 2007. MEDICAL EDUCATION 2007; 41: 1032–1038
between different types of skill and students with low interpersonal skills are more negative in their judgements. \(^{13,14}\) Therefore, the selection of appropriate students for peer evaluation is crucial. In our setting, the STs were very well aware of what the faculty expected from the students because they had actively participated in the teaching.

Students have mixed feelings about assessing each other. \(^{19}\) Asch et al. reported that students did not feel that they could objectively evaluate their peers during clerkships. \(^{20}\) In our study, most students felt they could rate younger students in an OSCE.

CONCLUSIONS

The results of our study show that adequate peer evaluation in OSCEs is possible and, consequently, most students both accepted the peer rating and felt confident that they could carry out peer ratings.

Our results are of importance for both small departments with limited staff that want to establish an OSCE and for larger departments that feel overwhelmed by the workload associated with running an OSCE.

As the OSCE at our medical school is a rather new and challenging form of assessment, it is important that it is accepted not only by faculty, but also by medical students. Our study demonstrated that the contribution of peers to the rating procedure met with student approval.

Future studies should investigate which skills are suitable for peer evaluation and which conditions favour reliable rating of skills by peers in an OSCE.

Contributors: J-FC, AS-N, TF and MMK contributed to the main development of the project. AK, MS, BE and BS supported the implementation, training of raters and development of OSCE stations. J-FC, AK and WH developed the questionnaire. AS-N and BS trained the standardised patients. J-FC performed statistical analysis, with support from AK and WH. J-FC and WH wrote the first draft; all authors revised the paper.

Acknowledgements: We would like to thank all general practitioners who participated as raters in the objective structured clinical examination.

Funding: We received an internal grant from the University of Göttingen Medical School to establish the objective structured clinical examination.

Conflicts of interest: None.

Ethical approval: Not required.

REFERENCES


Received 4 July 2006; editorial comments to authors 4 September 2006, 1 December 2006, 15 May 2007; accepted for publication 15 June 2007