The use of the objective structured clinical examination (OSCE) in dental education
Schoonheim-Klein, M.E.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 5

On the Reliability of a dental OSCE, using SEM: effect of different days.

M. Schoonheim-Klein¹, A. Muijtjens², L. Habets¹, M. Manogue³, C. van der Vleuten², J. Hoogstraten¹, U. van der Velden¹.

¹Academic Centre for Dentistry Amsterdam (ACTA), Amsterdam, the Netherlands.
²University of Maastricht, Maastricht, the Netherlands.
³Leeds Dental Institute, Leeds, UK.

European Journal of Dental Education, 2008; accepted for publication
Abstract

Aim: The first aim was to study the reliability of a dental objective structured clinical examination (OSCE) administered over multiple days, and the second was to assess the number of test stations required for a sufficiently reliable decision in three score interpretation perspectives of a dental OSCE administered over multiple days.

Materials and methods: In four OSCE administrations, 463 students of the year 2005 and 2006 took the summative OSCE after a dental course in comprehensive dentistry. The OSCE had 16–18 5-min stations (scores 1–10), and was administered per OSCE on four different days of 1 week. ANOVA was used to test for examinee performance variation across days. Generalizability theory was used for reliability analyses. Reliability was studied from three interpretation perspectives: for relative (norm) decisions, for absolute (domain) and pass–fail (mastery) decisions. As an indicator of reproducibility of test scores in this dental OSCE, the standard error of measurement (SEM) was used. The benchmark of SEM was set at <0.51. This is corresponding to a 95% confidence interval (CI) of <1 on the original scoring scale that ranged from 1 to 10.

Results: Mean weighted total OSCE score was 7.14 on a 10-point scale. With the pass–fail score set at 6.2 for the four OSCEs, 90% of the 463 students passed. There was no significant increase in scores over the different days the OSCE was administered. ‘Wished’ variance owing to students was 6.3%. Variance owing to interaction between student and stations and residual error was 66.3%, more than two times larger than variance owing to stations’ difficulty (27.4%). The SEM norm was 0.42 with a CI of ±0.83 and the SEM domain was 0.50, with a CI of ±0.98. In order to make reliable relative decisions (SEM <0.51), the use of minimal 12 stations is necessary, and for reliable absolute and pass–fail decisions, the use of minimal 17 stations is necessary in this dental OSCE.

Conclusions: It appeared reliable, when testing large numbers of students, to administer the OSCE on different days. In order to make reliable decisions for this dental OSCE, minimum 17 stations are needed. Clearly, wide sampling of stations is at the heart of obtaining reliable scores in OSCE, also in dental education.
Reliability of a dental OSCE

Introduction

Decisions for readiness of dental students to provide comprehensive care to patients should be based on examinations that are both valid and reliable and are based on accurate, appropriate, objective and unbiased information. Allowing incompetent candidates to practice, the false-positive decision is more damaging to patients than decisions to deny licenses to competent candidates, the false-negative decision. Re-examinations may correct these false-negative decisions, but it is never possible to correct the false-positive decisions (1).

One-time observations of clinical performance are highly unreliable (1). Unreliability may be attributed not only to examiner variation, but also to other uncontrolled sources of variability, or error (1). The main problem undermining reliability in the traditional clinical assessment is related to the biases introduced by examiners caused by the lack of standardisation of the tasks and criteria. To solve these problems in medical education the objective structured clinical examination (OSCE) has been introduced by Harden (2) in 1979. An OSCE is a clinical competency examination, where the student has to rotate between test stations (n = 10 to 20) with a clinical assignment, of mostly 5 or 10 min duration. In every station, the student's performance will be observed and assessed by an examiner, using a multi-item criteria checklist. The original development of OSCE anticipated that the standardisation of the task and the scoring (based on checklist-type rating forms) would enhance inter-examiner reliability and would solve the problems of reliability in traditional examinations (2). The OSCE did indeed improve inter-examiner reliability (3, 4), but later research in medical education concluded that the greatest threat to reliable measurement in performance examinations is case specificity (5–7). Case specificity involves the variation in performance of each student from one station to the other. To achieve an adequate level of inter-case reliability, many cases or stations are required (5, 8) to identify expertise. Thus, the unit of reliability analysis must necessarily be the station in the OSCE.

Other examples of sources of error, bias and thus threats to reliability in OSCE may include patient effects (9), the length of a station and testing time, native language (10), gender (11), ethnicity (10, 12), effects of the order in which students take the stations, student's fatigue or stress (13, 14), and the context of the learning environment (9, 10, 15, 16).

All these measurement errors can be studied separately, but in doing so, the relative contribution of each source of error to the total error variance remains unclear. Therefore, complex performance assessments, like the OSCE, require a more complex reliability model, such as the Generalizability theory analysis to account for various sources of measurement error in complex designs and to estimate the consistency of the generalisations to a universe or domain of skills (17).

Furthermore, in dental education, clinical assessment methods proved unsatisfactory. Chambers (1) concluded that the overall reliability of all one-shot initial dental licensure examinations given in the United States is approximately 0.40 and Manogue et al. concluded that, in the United Kingdom, results of clinic assessment were unsatisfactory (18). Therefore, the OSCE was also implemented in dental education (19–21). The reliability of dental OSCE was studied by Brown et al. (22) using a classical reliability measure: Cronbach's alpha* (alpha was 0.68). Gerrow et al. (23) preferred

* A glossary of terms used in studies of reliability is appended
reliability as a measurement of internal consistency, the Kuder–Richardson formula 20 (KR20).
*The KR20 was 0.69–0.74, indicating reasonable reliability.

Regarding the literature, there is to our knowledge no study carried out on the reliability of a dental OSCE using Generalizability theory and multiple indices of reliability. Reproducibility of scores of the OSCE can be studied from three test-score interpretation perspectives: for relative, absolute and pass–fail decisions.
1. **Norm test perspectives** for **relative** decisions: how is the student performing in comparing with the group?
2. **Domain test perspective** for **absolute** decisions: how is the student performing in absolute terms; how much of the domain is mastered?
3. **Mastery test perspective** for **pass–fail** decisions: will the student's score be above or below the pass–fail cut-off score (regardless how much above or below)?

As our dental school has a large intake with 158 students a year, there are more sessions on different days with multiple test forms of the OSCE in 1 week. Therefore, there is a need to test the differences in mean scores caused by the different days. The hypothesis is that the mean score of the students on day 1, Monday, is lower than that for subsequent days.

The aim of this study is first to study the effect of more sessions on different days of the OSCE in 1 week on the student's performance and second to assess the number of stations required for a sufficiently reliable decision in different perspectives in a dental OSCE.

**Materials and Methods**

Four consecutive dental OSCE were administered as summative assessment of the course in comprehensive care in a Training Group Practice setting: in June 2005 (n = 152, 18 stations), October 2005 (n = 106, 17 stations), January 2006 (n = 127, 17 stations), and June 2006 (n = 78, 16 stations), respectively. Each OSCE was administered on four different days of 1 week (Monday, Wednesday, Thursday and Friday mornings), using four slightly different test forms, during a 1-week period for all third-year students. There were two consecutive, early and late morning sessions, each consisting of the same stations and lasting 90 min. The four test forms varied only in the assignment of examiners and minor variations, like a different X ray or different teeth, not interfering with the criteria. Students were randomly allocated to day of administration. In total, 463 students were assessed with a mean number of 17 stations. A blueprint (Table 1) was constructed using five clinical disciplines, Periodontology, Cariology, Oral Function, Orthodontics and Radiology. The OSCE tested six domains of competence: Diagnostics, Diagnostics on radiographs, Health promotion, Treatment, Practice management and Communication. The performance of the students was rated in every station by means of a checklist, with 10 items or criteria. For each correct item, 1 point was given and the total score per station thus varied between 1 and 10 points. For this study, the pass–fail cut score for the OSCE was estimated by the Borderline regression standard setting method (24). The computed pass–fail score of the OSCE was 6.2 on the scale of 1–10.
<table>
<thead>
<tr>
<th>Station Domain of Competence</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIAGNOSTICS</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIAGNOSTICS ON RADIOGRAPHS</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HEALTH PROMOTION</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TREATMENT</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRACTICE MANAGEMENT</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COMMUNICATION</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Chapter 5

Statistical analyses
Descriptive statistics were carried out using spss 11.5 (SPSS Inc., Chicago, Illinois, USA). Scores for each station are expressed on a 1–10 point scale (items correct, 1–10). Station scores were averaged across days and OSCE administrations were weighted by sample size of persons to calculate the overall weighted mean OSCE score. One-way ANOVA was used to test for differences between performances across days per OSCE administration. P values <0.05 were considered as statistical significant.

The GENOVA program (designed by Brennan, 1980) was used to perform a Generalizability analysis to estimate the reliability of the OSCE and to estimate the amount of stations required for a reliable pass–fail decision.

The reliability of a student's test score for the OSCE was estimated using a one facet* random effect, person (P) by station (S) separate for each of the days (P · S design). For each OSCE, per day variance components* were estimated for persons and stations and the interaction effect of person by station with residual effect (PS,e) were estimated. The resulting variance components were pooled across OSCE days and administrations were weighted by the sample size of persons to achieve one single average estimate.

Decision studies* were then conducted to estimate the required amount of stations, by estimating different levels of reliability coefficients: reproducibility of scores as function of the number of stations and thus the examination length.

Three test interpretation perspectives in order to determine the reliability
1. Norm test perspective for relative decisions: The reliability in this perspective was estimated with the Generalizability coefficient* (norm).
2. Domain test perspective for absolute decisions: The reliability in this perspective was estimated with the Dependability coefficient* (domain).
3. Mastery test perspective for pass–fail decisions: The reproducibility of the pass–fail decisions, the reliability, was established by estimating the adjusted Dependability coefficient (mastery), a reliability-like coefficient.

The customary benchmark for all three reliability coefficients is set at 0.80 (25).

These three reliability coefficients are dependent on the variability of the group of examinees. When there is no variance between examinees, there is no reliability. An alternative for expressing reliability in reliability coefficients is by the SEM*. SEM permits a calculation of the precision of the measurement of a score on the original scoring scale. Although this may hinder cross-study comparison, Norcini (25) explained that the SEM is a very intuitive way of expressing reliability and has at least two advantages over the reproducibility coefficients. It is less influenced by the variability among students, and the confidence intervals (CI) based on the SEM are easier to understand. Therefore, SEM is used in this study as the major reliability index.
Reliability of a dental OSCE

The SEM norm, in the relative test perspective, was estimated on the basis of the weighted mean variance components with the formula:

$$SEM_{norm} = \sqrt{\frac{PS}{N_s}}$$

where \(PS\) = the variance component of ‘interaction of persons with stations’ and \(N_s\) = number of stations.

For both the domain and mastery perspective, the corresponding SEM domain is the same. SEM domain as an indicator of reproducibility was estimated on the basis of the weighted mean variance components with the formula:

$$SEM_{domain} = \sqrt{\left(\frac{S}{N_s} + \frac{PS}{N_s}\right)}$$

where \(S\) = the variance component ‘stations’. The benchmark of both the SEM in the three perspectives was set at <0.51, in order to make a reliable inference of student’s scores within at least one unit on the 10-point scoring scale. This benchmark is corresponding to a 95% CI of <1 on the original scoring scale, that ranged from 1 to 10.

Results

OSCE performance across days and variance components

Table 2 shows the mean total checklist scores and standard deviation (SD) of the different OSCE on the four different days with the number of students and stations for each OSCE. The overall mean (weighted for persons) score of the four OSCE was 7.14. There is no systematic change or pattern over the 4 days. ANOVA showed no significant differences of the mean scores between the days; only in the OSCE of June 2005, the scores on Friday, the last day of the OSCE, were lower than the Thursday scores.

The variance components (and their standard errors) pooled, weighted for persons, across all OSCE administrations as results of the Generalizability study for the (P x S) design are shown in Table 3. The score variance was attributed to 6.3% students, 27.4% stations difficulty and 66.3% interaction and residual error.

Reliability indices as a function of stations and testing time

The Generalizability coefficient (norm) ranged from 0.52 to 0.68 and was 0.62 for the mean of four OSCE and 17 stations. The corresponding SEM norm is 0.42. An SEM norm of 0.42 provides an estimated 95% CI of an individual score between 0.82 points added to or subtracted from the individual score. Regarding the domain perspective, the Dependability coefficient (Phi) was 0.54 for the mean of four OSCE and 17 stations. Regarding the mastery perspective, the adjusted Dependability coefficient was 0.82 for the mean of four OSCE and 17 stations. The corresponding
**TABLE 2. Mean total checklist scores and standard deviations from sub-samples of four OSCEs and overall analyses**

<table>
<thead>
<tr>
<th>OSCE</th>
<th>Number of test-stations</th>
<th>Number of students</th>
<th>Test Form Mean score (sd)</th>
<th>Mean total score (n)</th>
<th>Students passed</th>
<th>Pass rate %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Day1</td>
<td>Day2</td>
<td>Day3</td>
<td>Day4</td>
</tr>
<tr>
<td>June 2005</td>
<td>18</td>
<td>152</td>
<td>7.19 (0.65)</td>
<td>7.03 (0.75)</td>
<td>7.29 (0.66)</td>
<td>6.79 (0.75)*</td>
</tr>
<tr>
<td>October 2005</td>
<td>17</td>
<td>106</td>
<td>6.55 (0.84)</td>
<td>7.11 (0.62)</td>
<td>6.90 (0.66)</td>
<td>7.08 (0.56)</td>
</tr>
<tr>
<td>January 2006</td>
<td>17</td>
<td>127</td>
<td>7.16 (0.92)</td>
<td>7.56 (0.69)</td>
<td>7.41 (0.69)</td>
<td>7.33 (0.61)</td>
</tr>
<tr>
<td>June 2006</td>
<td>16</td>
<td>78</td>
<td>7.16 (0.61)</td>
<td>7.03 (0.68)</td>
<td>7.35 (0.60)</td>
<td>6.99 (0.74)</td>
</tr>
<tr>
<td>Overall</td>
<td>Mean 17</td>
<td>Sum 463</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* p<0.05 Only in the OSCE June 2005, day 4 showed significant lower mean scores than day 3
SEM domain is the same for both absolute and pass–fail decisions. SEM domain is 0.50. An SEM domain of 0.50 provides an estimated 95% CI of an individual (pass–fail) score between 0.98 points added to or subtracted from the individual (pass–fail) score.

The required number of stations for reliable OSCE results

The results of the decision studies to predict the required number of stations and thus varying testing time estimates for reliable OSCE results are shown in Table 4.

The benchmark of the Generalizability coefficient (norm) is reached at 40 stations, the benchmark of the Dependability coefficient (domain) is reached at 60 stations and the adjusted Dependability coefficient (mastery) is reached at 15 stations.

The predicted corresponding SEM norm and SEM domain and 95% CI for different numbers of stations are shown in Table 5 and Table 6, respectively. The minimum amount of stations required in relative interpretation perspective using SEM norm <0.51 as criterion for reliability is 12 stations. The minimum amount of stations required in absolute and mastery perspective, using SEM domain <0.51 as criterion for reliability is 17 stations.

Discussion

There was no significant improvement of scores during the OSCE week. In one OSCE, it was even reversed: the scores on Friday were lower than the scores on the earlier days in the week. If information was shared by the students, it had no systematic effect on scores. This is in agreement with the results of studies conducted by Rutula et al. showing a general lack of significant improvement in the scores over the days of testing, as in some other examinations (26, 27). The present findings support the suggestion that even if the transmission of information among students occurred, it has a minimal effect on the scores of the students.

In a mastery-oriented perspective, the reproducibility of pass–fail decisions is estimated, rather than the reproducibility of scores. Using a mastery-oriented test perspective on score interpretation ameliorates the reproducibility of OSCE (8, 25, 28). In this respect, the results of the present study are encouraging, because in the mastery test perspective, 15 stations are required for dental OSCE.
TABLE 4. Generalizability coefficients, Decision-coefficients for checklist scores and adjusted Dependability coefficient for pass-fail cut score 6.2, as a function of testing time.

<table>
<thead>
<tr>
<th>OSCE administration</th>
<th>G-coefficient (norm perspective) as function of examination-time in hrs</th>
<th>D-coefficient (domain perspective) as function of examination-time in hrs</th>
<th>Adjusted D-coefficient (mastery perspective) Pass-fail cut score 6.2</th>
</tr>
</thead>
<tbody>
<tr>
<td># Hrs of testing</td>
<td>1.5hrs</td>
<td>1.7hrs</td>
<td>2.5hrs</td>
</tr>
<tr>
<td>N test-stations</td>
<td>18</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>June 2005</td>
<td>0.66</td>
<td>0.68</td>
<td>0.77</td>
</tr>
<tr>
<td>October 2005</td>
<td>0.53</td>
<td>0.56</td>
<td>0.66</td>
</tr>
<tr>
<td>January 2006</td>
<td>0.69</td>
<td>0.71</td>
<td>0.79</td>
</tr>
<tr>
<td>June 2006</td>
<td>0.61</td>
<td>0.64</td>
<td>0.72</td>
</tr>
<tr>
<td>Overall OSCE</td>
<td>0.64</td>
<td>0.66</td>
<td>0.74</td>
</tr>
</tbody>
</table>

*5 minutes per station.

TABLE 5. Standard Error of Measurement (SEM-norm) and confidence intervals (CI) in relative interpretation perspective as a function of examination time and number of stations. (For the design (P x S))

<table>
<thead>
<tr>
<th>OSCE year</th>
<th>SEM G (norm perspective)</th>
<th>95% Confidence Interval (SEM G*1.96) (10-point scale)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sq rt (PS/Ns)</td>
<td>1,5 hrs (18 st) 1,7 hrs (20 st) 2,5 hrs (30 st) 3,3 hrs (40 st) 1 hr (12 st) 1,5 hrs (18 st) 1,7 hrs (20 st) 2,5 hrs (30 st) 3,3 hrs (40 st) 1 hr (12 st)</td>
</tr>
<tr>
<td>June 2005</td>
<td>N stations 18</td>
<td>N students 152</td>
</tr>
<tr>
<td>October 2005</td>
<td>17</td>
<td>106</td>
</tr>
<tr>
<td>January 2006</td>
<td>17</td>
<td>127</td>
</tr>
<tr>
<td>June 2006</td>
<td>16</td>
<td>78</td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
<td>463</td>
</tr>
</tbody>
</table>

Bench mark

Bench mark

Bench mark

Bench mark
TABLE 6. Standard error of measurements (SEM-domain) and Confidence Interval (CI) in absolute interpretation perspective, as a function of examination time and number of stations. (For the design (P x S)).

| OSCE year | N stations | N students | 1.5 hrs (18 st) | 1.7 hrs (20 st) | 2.5 hrs (30 st) | 3.3 hrs (40 st) | 1.4 hrs (17 st) | 1.5 hrs (18 st) | 1.7 hrs (20 st) | 2.5 hrs (30 st) | 3.3 hrs (40 st) | 1.4 hrs (17 st) |
|-----------|------------|------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| June 2005 | 18         | 152        | 0.48           | 0.46           | 0.37           | 0.32           | 0.50           | ±0.94         | ±0.90           | ±0.73           | ±0.63           | ±0.98           |
| October 2005 | 17     | 106        | 0.52           | 0.50           | 0.40           | 0.35           | 0.54           | ±1.02         | ±0.98           | ±0.78           | ±0.69           | ±1.06           |
| January 2006 | 17    | 127        | 0.47           | 0.45           | 0.37           | 0.32           | 0.49           | ±0.92         | ±0.88           | ±0.73           | ±0.63           | ±0.96           |
| June 2006 | 16         | 78         | 0.47           | 0.45           | 0.36           | 0.31           | 0.48           | ±0.92         | ±0.88           | ±0.71           | ±0.61           | ±0.94           |
| Total     | 17         | 463        | 0.49           | 0.46           | 0.38           | 0.33           | 0.50           | ±0.96         | ±0.90           | ±0.74           | ±0.65           | ±0.98           |

(For the design (P x S)).
in general with the benchmark of adjusted D coefficient of 0.80 (Table 4). Even more encouraging is the related SEM domain for this OSCE, which required only 17 stations. The SEM domain was 0.50 with a CI of 0.98. The CI of 0.98 of the pass–fail score 6.2 gives a range of the pass–fail score between 5.22 and 7.18. To make reliable decisions about the passing and failing of students, the combination of the adjusted D coefficient and the SEM in finding an optimum amount of stations seems satisfactory.

The tolerance for error depends on the context and intended use of the results of measurement, as stated by Kane (1996). A standard error of an ounce in a bathroom scale represents unnecessary precision. The same standard error in a laboratory scale used to weigh tissue samples would be unacceptably large (29). In this OSCE, the tolerance of SEM on the score 1–10 is <1. The weighted mean score of four OSCE was 7.14, with the SEM domain of 0.50, the 95% CI for a student checklist scores on this OSCE is ±0.98. (i.e. precise enough to distinguish within one unit of the scoring scale). Perhaps, in very high stake examinations, this might still be too imprecise and more stations need to be sampled. To reduce the measurement error to 0.33, approximately the double number of stations is required (> 40 stations).

The total variance component for persons (students) pooled across the four OSCEs accounts for 6.3% of the total variation in the checklist scores. Although it seems a small percentage of the total variance, the dental OSCE is discriminating between students comparable with other studies, where the variance for persons accounted for 3% and 11% of the total variance (7, 8). Perhaps, the use of global rating scales would better discriminate between persons, as was concluded by Govaerts (8). This looks like a contradiction to the reason why originally the OSCE was developed, i.e. because of the lack of standardisation. However, the classic, global, clinical assessment was apparently less reliable not because of the lack of standardisation and structuring, but the lack of adequate sampling (30). The results in the studies reviewed by Williams et al. for not standardised and unstructured global clinical performance evaluations suggest that a minimum of 7 to 11 clinical examiner ratings will be required to estimate reliable overall clinical competence (31). This will be explored in further research.

The variance component reflecting the student by station interaction and general error (PS,e) is the largest source of variance: 66.3% of the variance of the checklist scores (two times larger than variance owing to station difficulty 27.4%). Thus, some students are good in diagnostics and others are better in treatment. In conclusion, the problem of case specificity is also present in dental OSCE, and this confirmed other studies (7, 8). Various strategies have been adopted to minimise the practical difficulties raised by case specificity. The simplest is to combine the OSCE with other test formats that provide more efficient sampling of content. As long as all examination components are based on the same blueprint, this is a justifiable approach (32, 33). Therefore, it is suggested that a written component of post-test OSCE stations with questions about the preceding test stations would enhance the reliability.

There are several ways to improve the reliability of assessments. Most important is the use of sufficiently large numbers of cases (stations) to adequately sample the domain of interest. It is also advised to use examination questions or performance cases that are of medium difficulty for the
students being assessed. If stations had a unilateral outcome, when nearly all students get most questions correct or incorrect, little information is gained about student achievement, and the reliability of these stations would be low. This happened in two stations in this study about the radiology diagnostics. The station's design and the education about these subjects need further improvement in order to ameliorate the reliability of the OSCE scores.

**Conclusion**

The present dental OSCE showed no significant improvement in the scores over the days of testing. It appeared safe when testing large numbers of students to administer the OSCE on different days. In order to make reliable decisions about passing or failing this dental OSCE, minimum 17 test stations are needed and more in very high-stake circumstances. Clearly, wide sampling of stations is at the heart of obtaining reliable scores in OSCE, also in dental education.

**Acknowledgement**

The authors would like to thank Ron J. I. Hoogenboom, research assistant at the Department of Educational Development and Research, University of Maastricht, for his assistance in data analyses.
Glossary

The terms used by various authors for different theories of reliability vary; hence, it is important to ascertain which working definitions are being used. Reliability in this article is based on the notion of the reproducibility of scores.

*Adjusted Dependability coefficient* is reliability-like coefficient, in a mastery-oriented interpretation perspective, and gives an estimate of the reproducibility of the pass–fail decisions.

*Classical Test Theory* is a special case of Generalizability theory in which error is unitary.

*Confidence intervals* around a particular (cut-off) score owing to unreliable variance can be estimated with 68% probability if one SEM plus or minus is used, or with 95% if two SEM plus or minus are used.

*Cronbach’s alpha* is a reliability coefficient derived from classical test theory, which measures the degree of internal consistency of a test, providing a score from 0 to 1.

*Decision (D) studies* emphasise the estimation, use and interpretation of variance components for decision making, to estimate reliability coefficients under different facets, e.g. the impact of varying number of cases and judges on the reproducibility of scores.

*Dependability (D) coefficient* (Phi) is a reliability-like coefficient in the absolute referenced interpretation perspective. The D coefficient is always smaller than Cronbach’s alpha, because it takes into account all other possible sources of error that affects not only the rank ordering but also the absolute scores.

*Facet* is a set of similar measurement conditions; e.g. stations, judges and time of day.

*The Generalizability (G) study* is conducted to estimate the sources of variance components associated with a universe of admissible observations. The G study asks to what extent a sample of measurement generalises to a universe of measurement.

*Generalizability theory* is a statistical theory of measurement, developed by Cronbach et al. (34) that expands classical test theory to include all sources of measurement error variance simultaneously.

*Generalizability (G) coefficient* is a reliability coefficient in the norm-referenced interpretation perspective and is an intra-class correlation comparable with Cronbach’s alpha, providing a score from 0 to 1.

*KR20* is similar to Cronbach’s alpha but usually used for dichotomous scoring. KR20 and Cronbach’s alpha are intended for use with norm-referenced score interpretation.

*Standard error of measurement (SEM)* is an index of the reproducibility expressed on the same scale of the scores in a domain or in a norm interpretation perspective. SEM permits a calculation of the precision of measurement at the (cut) score. A low SEM means a high reliability.

*Variance components* are the estimated sources of variance from a G study computed by random effect one-way analyses of variance.
References