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

Citation for published version (APA):
Chapter 6

Who will pass the dental OSCE?
Comparison of the Angoff and the Borderline Regression standard setting methods

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

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

Submitted for publication in the European Journal of Dental Education
Chapter 6

Abstract
Aim of this study is to elucidate which standard setting method is the optimal instrument to prevent that incompetent students pass and competent students fail a dental OSCE.

Material and methods: An OSCE with 14 test-stations was used to assess the performance of 119 dental 3rd year students in a training group practice. Three standard setting methods were applied: the Angoff I method, the modified Angoff II with reality check and the Borderline Regression (BR) method to establish the pass/fail standard per station. For the final decision about passing or failing the complete OSCE, the 3 methods were compared in a Total Compensatory (TC), a Partial Compensatory (PC) within clusters of competence, and a Non-Compensatory (NC) model.

The reliability of the pass/fail standard of the 3 methods was indicated by the Root Mean Square Error (RMSE). As a criterion measure, a sample of the students (n=89) was rated in the clinic by their instructors and accordingly these students were divided into two groups: competent and incompetent students. The students’ clinical rating (considered for this study as “true qualification”) was compared with the pass-fail classification resulting from the OSCE. Undeserved passing of an incompetent student was considered as more damaging than failing of a competent student.

Results: The BR method showed more acceptable results than the two Angoff methods.

In terms of pass rate the BR method showed the highest pass rates: for the TC model the Angoff method I and II and the BR showed the pass rates: 86.6%, 86.6%, and 97.5 % respectively. For the PC model the pass rates were 30.3%, 34.5%, and 61.3%, and for the NC model the pass rates were 0.8%, 1.7% and 7.6%. The BR method showed lower RMSEs (higher reliability): for the TC model the RMSEs were 1.3%, 1.0% and 0.3% for the Angoff I, Angoff II and BR method respectively, and for the PC model, the RMSE of the clusters of competence ranged for Angoffs I: 2.0%-3.7%; for Angoff II: 1.8%-2.2%, and for the BR method 0.6%-0.7 %. In terms of incorrect decisions, the BR method had a higher loss due to incorrect decisions for the TC model than for the PC model.

Conclusion the Borderline Regression method applied for a partial compensatory model provides defensible pass/fail standards and seems to be the optimal choice for this type of dental OSCE.
Introduction

The examinations in which students have to perform and the decisions derived from these must be defensible to the examinees, the examiners, the organisation and to society. Stated differently (1), competent students should pass and incompetent students should fail an OSCE. Since 1979 the OSCE is widely used for testing competencies in medical education (2). Around 1997, the OSCE was also introduced in dental education (3-5). An OSCE is a clinical competency test where the student rotates between 10-20 test-stations with an assignment of mostly 5 or 10 min duration. In each station the student’s performance is observed and assessed by an examiner, using a multi-item criteria checklist.

In a clinical examination, like an OSCE, a cut-off score or a pass/fail standard is a point on the observed score scale that serves as a boundary to divide between students who performed satisfactory and those who did not (6). The pass/fail standard is a conceptual boundary on the “true-score scale”, and indicates the minimal adequate level of competence (7). Standard setting is a judgemental process, often with arbitrary decisions about what is “good enough”. The use of standard setting methods helps to ensure that decisions or classifications are based on non arbitrary explicit criteria, which are combined in a systematic, reproducible, objective, and defensible manner (7-11). This procedure will result in a defensible pass/fail standard.

Standards can be categorized as either relative (norm referenced) or absolute (criterion referenced) (9, 12). Relative standards are mostly used for admission purposes, and are expressed as a fixed percentage of examinees that should pass, regardless their absolute performance, e.g. the pass/fail score is set at the score that passes 60% of the group. Relative standards vary with the performance-level of the group. Standards are absolute (criterion-referenced) when the pass/fail point is set on the basis of what the student should know or be able to do. Absolute standard setting is the preferred option for a test of clinical competence in for example an OSCE (8, 12). Within the absolute standard setting methods, Jaeger (13) made a fundamental distinction between rational (test-centred) methods and empirical (examinee-centred) methods. In test-centred methods the content of the test is analysed or reviewed by judges in order to compute a standard; e.g. the Angoff standard setting method (14, 15). In the Angoff procedure, judges are asked to review each item in a test and to estimate “the probability that the ‘minimally acceptable candidate’ would answer (or do) an item correctly” (Angoff, 1971, p. 515). Often judges tend to set the Angoff standard too high. Therefore in a modified Angoff method judges are allowed to adjust their estimates after being provided with true performance scores of candidates (a reality check). On the other hand, in an examinee-centred method the students performance is observed by experienced practitioners and scored on a checklist and additionally also rated globally on a Likert scale as incompetent, borderline, sufficient, good or excellent, and this information is used to set the pass/fail standard, e.g. the Borderline Group method where the pass/fail standard is determined by calculating the mean checklist score for the subgroup of examinees rated as borderline (1, 16, 17). Another example is the Borderline regression method (BR): this method uses the data of the complete group of examinees and uses linear regression of the checklist scores onto the global ratings to produce a linear equation. By inserting the point of the borderline score of the global rating scale
into the equation, a corresponding predicted checklist score can be determined. This predicted score becomes the pass/fail score for the checklist score of the station (18).

Research on standard setting indicates that different methods for setting standards may lead to different results (13). Standard setting has been called the “Achilles Heel” of educational testing (19) primarily because there is no clear consensus on the best choices among numerous methods and because the results of applying any method cannot easily be validated (20). As Kane (20) stated, the validation of a performance standard can never be completely tested. The decision about passing competent students and failing incompetent students is based on the chosen pass/fail score from a specific method. The validation of the decision process in which a pass/fail standard is used, does not constitute a proof that the pass/fail standard is correct or true, but it supports the credibility and defensibility of the chosen pass/fail score. (21). Therefore instead of validation it is more appropriate to collect evidence to support the credibility of the standards (22). The credibility, the evaluation of consequences of the standard, could be estimated by comparing the pass rates against contemporary markers/levels of competence (9, 22, 23).

In a review on standard setting in medical education Cusinano (7) described many studies about standard setting methods also for OSCEs. This evaluation showed great variability in standards from different methods. Therefore he stated that research is needed to study reliability of well-known standard setting methods and apply them to formats such as the OSCE.

Both standard setting methods, Angoff and BR, have been applied to OSCEs (1, 15, 16, 18, 23, 24). Yet, recent studies show inconsistent results, when comparing the reliability and credibility of the Angoff and the BR method applied in OSCEs. One study showed that the BR method provides a more reliable and credible standard (23), but another study concluded that BR and Angoff are both credible and the reliability could not be compared (24). To our knowledge there are no studies about standard setting in dental education.

**Scoring models**

A test scoring model refers to the way in which station-scores are combined to arrive at a total result of the test (11). Many experts recommend a total compensatory approach when performing standard setting for an OSCE (23, 25). The question is whether the final pass/fail decision should be based on the overall score across all stations, in a fully compensatory model, or on a model without compensation between stations (7, 11, 26). Another possibility may be a combination of these in a mixed model. Such a model ensures that students reach a minimum level of competence in special domains of competence, like communication or diagnostics (10, 27). Newble (26) stated no approach is more valid than others and Norcini (6, 9) recommended for future research to study the effects of mixed compensatory models. Therefore it can be questioned, which compensatory model is most optimal for dental OSCEs: 1) the total compensatory (TC) model where the average of all station scores is compared with the average of all station standards to reach a pass-fail decision, 2) the partial compensatory (PC) model where each station is assigned to a cluster, a pass–fail decision is reached per cluster by comparing the cluster’s average score and average standard, and a pass for the OSCE is only reached when all clusters are passed, and 3) the non-compensatory
model (NC) requiring all stations to be passed in order to reach a pass for the OSCE.

The aim of the present study was to elucidate which standard setting method, including different compensatory models, is the optimal instrument to prevent that incompetent students pass and competent students fail a dental OSCE. The standard setting methods tested in this study are the Angoff, the modified Angoff with a reality check, and the Borderline Regression standard setting method.

**Material and Methods**

**Sample and design of the study**
In this study an OSCE with 14 test-stations was used to assess 119 dental students. Three standard setting methods were applied to establish the pass/fail standard per station: the Angoff I method, the modified Angoff II with reality check and the Borderline Regression (BR) method. The reliability of the pass/fail scores and the credibility of the methods were investigated for the three compensatory models presented in the introduction. Finally, it was determined, which combination of standard setting method and compensatory model is most appropriate for a dental OSCE.

**Overall performance in the Dental Training Practice**
Third year students in the Bachelor phase of dental education in ACTA are participating one day a week in a comprehensive dentistry course within a dental training practice setting. In order to treat more complex patients in the Master phase of the study, the requirements for a positive judgement of their performance in the training practice are threefold: a sufficient OSCE score, a fixed amount of procedures (“fillings”), and a qualitative assessment of their way of handling patients. During the comprehensive dentistry course, the dental instructors rate their students’ performance with patients on a regular basis for feedback. For this study, these instructors were asked to rate the overall performance in the comprehensive dentistry course of a sample of their students (n=89) on a 5 point scale: incompetent, borderline, sufficient, good, and excellent. These students were divided according the overall performance score in two groups: the clinically incompetent students rated as “poor” and “borderline”, and the clinically competent students: rated as “sufficient” or “good” or “excellent”.

**OSCE**
After completion of the comprehensive dentistry course, all students (n=119) were assessed by a summative 14-station OSCE. The OSCE covered 4 competency domains, based on the profiles of the European dentist: I Professionalism, II Communication and interpersonal skills, III Clinical information gathering and Diagnosis and Treatment planning, IV Establishment and maintenance of oral health. These domains were assessed in 4 clusters with 3 or 4 stations per cluster. The stations were observed and rated by staff (examiners) using checklists that required yes/no responses on 10 items of adequate performance, yielding test scores per station ranging from 0-10. For standard setting purposes, additionally and independent from the item checklist...
MARKING resuscitating Annie

- Check it is safe to approach the patient 1 0
- Attempt to arouse the patient 1 0
- Place chair horizontal 1 0
- Tilt head backwards and check the airway 1 0
- Establish whether the patient is breathing for 10 seconds 1 0
- Call for help 1 0
- Place patient on the floor 1 0
- Undertake cardiopulmonary massage 30 compressions 1 0
- Inflate the lungs twice to correct extent 1 0
- Undertaken in right order 1 0

Assessor: ____________________ Total Marks out of 10: ____________________

Global rating: This student is: 1 □ poor 2 □ borderline 3 □ sufficient 4 □ good 5 □ excellent

Fig. 1. Example of marking scheme to assess a student in an OSCE station using 2 scoring methods: scoring on a checklist (0-10) and for the Borderline Regression method scoring on a global rating list (1-5).

scoring, each examiner rated the students’ performance in the OSCE station on a 5-point-global rating scale: 1.fail, 2.borderline, 3.sufficient, 4.good, and 5.excellent (see Fig.1). In this way in each station the examiner gave two scores: 1. the checklist score 0-10 and the global rating score 1-5.

In order to cope with the large number of students (n=119) six sessions of the OSCE were scheduled during 4 days: two days with two sessions and two days with one session per day. Each day the stations had the same format with small variations in the content e.g.: a different radiograph, a different story for the standardized patient, and different teeth to measure pockets. Some stations had the same examiner for all sessions; most stations had a different examiner on the various days.

Forty-two examiners were involved, all dental instructors with former OSCE experience. The OSCE consisted of 12 stations with one examiner who directly observed the student and 2 stations with an examiner and a simulated patient. The simulated patient and the examiner decided together on the scoring.

Standard setting methods
Three standard setting methods were used for this study to establish the 3 pass/fail standards of the stations: the Angoff I, the Angoff II, and the BR method.

A student obtains a pass if his/her score exceeds, or is equal to the pass/fail standard, and otherwise obtains a fail. The percentage of students passing the total OSCE accordingly is indicated as the pass rate. The pass rate will increase when a less stringent method is used.

The Angoff method
The Angoff standard setting method allows determining the pass/fail standard before the OSCE is actually taken. In our setting, thirteen examiners, with former OSCE-examiner experience, were asked to be a judge in the Angoff I procedure. The procedure started with a group discussion about the definition of the borderline student. The judges decided to consider a borderline student as a
student who is performing at the level between a clear pass or clear fail of the OSCE station. Then
the judgements about the items were made individually using detailed written instructions. When
a judge did not feel an expert in the subject of station he could resign from giving a judgement for
that station. The judges answered for each checklist-item in each station the question: “Yes or no,
will the borderline examinee respond correctly to this item?” According to Downing et al.(10),
this is the original Angoff method. Reported drawback of the Angoff method is that judges tend
to set the score too high (21). Therefore a reality check was implemented and the procedure was
repeated by the same judges. The judges could revise their first judgements after the OSCE. They
were provided with information of the station's mean performance score, pass/fail standards and
pass rates (Angoff II). For efficiency reasons, no opportunity for discussions after the ratings was
included in the Angoff procedures. For each station the Angoff I/II pass/fail estimate was obtained
for each judge by calculating the number of items (0-10) that were judged to be correctly answered
by a borderline examinee. Then by averaging these values over all judges the station's pass/fail
standard was obtained.

**The Borderline Regression (BR) method**

In the BR method in each station the students were rated on a 10-item checklist (0-10) and also for
this method on a 5-point-*global* rating scale (1-5) (see example in Fig.1). Per station the checklist
scores and the global ratings were compared with linear regression. The BR-pass/fail standard per
station was obtained by using the regression line to calculate the checklist score corresponding
with the cut-off point “2” (borderline) of the global rating.

**Compensatory models**

Using the pass/fail standards per station, the total test result for the OSCE was calculated for the
three compensatory models presented in the Introduction section: a conjunctive model with No
Compensation (NC) between stations, a mixed model with Partial Compensation (PC) between
stations, and a Total Compensatory model (TC), with compensation between all stations.

**Reliability of 3 pass/fail standards (Angoff I, Angoff II and BR) within 3 compensatory
models**

The reliability of the pass/fail standards of the clusters (PC) and the total OSCE (TC) was
investigated on the basis of the root mean square error (RMSE) for each standard setting method.
The RMSE is an estimate of the standard error of the pass/fail standard obtained in the standard
setting methods.
The reliability of the Angoff method was assessed using generalisability theory by estimating the RMSE of the pass/fail standard of each cluster according to the following formula:

$$RMSE_{\text{Angoff}} = \sqrt{\frac{\sigma_{js}^2}{\sum N_{js}}}$$

where $\sigma_{js}^2$ is the estimated variance component for the effect of judges nested in stations, the error, and $N_{js}$ is the number of judges providing Angoff estimates for the standard of the station $s$.

For the BR method the RMSE of the pass/fail standard of a cluster was defined as

$$RMSE_{\text{BR}} = \sqrt{\frac{1}{M^2} \cdot \frac{1}{n} \cdot \sum \left( \bar{s}_{\text{reg},i}^2 \cdot (1 + \frac{(R_o - MN_{R,i})^2}{[(n_o - 1)/n_o] \cdot SD_{R,i}^2}) \right)}$$

where $M$ is the number of stations in the cluster, $n_o$ the number of students that attended the stations of the cluster, $s_{\text{reg},i}^2$ is the standard error of estimate of the regression of the check list score on the global rating, and $MN_{R,i}$ and $SD_{R,i}$ the mean and standard deviation of the global score for the $i$-th station in the cluster, respectively, $R_o$ the pass/fail cutoff for the global rating, and $n$ the (hypothetical) number of students the $RMSE_{\text{BR}}$ is estimated for.

**Credibility of the 3 pass/fail standards for 3 compensatory models**

Comparable to the evaluation of the diagnostic performance of biomedical tests by measuring the sensitivity and specificity, also the decisions about failing or passing the OSCE based on the standard have these two characteristics associated with the two types of classification errors: false negatives and false positives (30). Sensitivity is the true positive rate, the proportion of competent students, which are correctly identified by the test, i.e. by an OSCE. Specificity is the true negative rate, the proportion of incompetent students, which are correctly identified by the test. To establish the credibility of the 3 standard setting methods in 3 compensatory models the following question needs to be answered for each standard setting method: Which proportion of students passed the OSCE, while they should have failed according to the overall clinical rating and which students failed while they should have passed? Table 1 provides the nomenclature of these decisions.

| TABLE 1. Nomenclature, four possible coincidences of decisions in diagnostic performance of standard setting methods with different compensatory approaches in incompetent and competent students for a dental OSCE |
|---|---|---|---|
| Overall clinical rating | Test (OSCE) exam decision with standard setting | Pass | Fail |
| Competent | TP | FN | TP+FN |
| Incompetent | FP | TN | FP+TN |
| Total | TP+FP | FN+TN | TP+FP+FN+TN |

TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative
Sensitivity = True Positive Rate = TP/ (TP+FN)
Specificity = True Negative Rate = TN/ (FP+TN)
In this study the Loss function was introduced as new criterion for credibility of the standard setting method. The Loss function provides a tool for evaluating the classification accuracy of diagnostic tests (31-33). The Loss function is the weighted sum of the sensitivity and specificity. Loss function analyses the numbers of incorrect decisions, i.e. the percentage of incompetent students, who pass (False Positive Rate, FPR), and the percentage of competent students who fail (False Negative Rate). More weight was assigned to the FPR, because the consequences to patients of passing an incompetent dental student for a clinical examination like an OSCE outweigh the consequences of failing a competent student. Reexamination can correct false-negative decisions (competent students failed), but not false-positive decisions (incompetent students passed) (34). Thus in other words: a high specificity is considered more important than a high sensitivity. Therefore the Loss function is analyzed for the weight of a false negative equal to one, and the weight of a false positive varying from 1 to 50. The standard setting method with the minimum weighted loss is considered as the most credible method.

Weighted loss was calculated for the 6 combinations of the 3 standard setting methods and 3 compensatory models with the formula:

\[
\text{Weighted Loss } (V) = \frac{\left[ PC \cdot (1 - TPR) + (1 - PC) \cdot V \cdot FPR \right]}{(1 + V)}
\]

Where PC is the proportion of Competent students (according to the dichotomized global clinical rating), TPR is the True Positive Rate and FPR is the False Positive Rate. V is the loss associated with a false positive classification, while the loss associated with a false negative is set to 1.

Finally by comparing the reliability and the credibility, it was determined which combination of standard setting methods and compensatory models was most appropriate.

**Statistical analyses**

For the descriptive analyses SPSS 11.5 (SPSS Inc., Chicago, Illinois, USA) was used. Mean (SD) OSCE checklist and global rating scores were calculated. Mean Angoff I, mean Angoff II, and mean BR pass/fail standards were calculated per station, per cluster of stations, and for the complete test. Differences between pass/fail standards of the three methods across stations were tested using a paired t-test, considering p<0.05 as statistically significant. Pass rates for the OSCE were calculated for each combination of standard setting method and compensatory model.

The reliability for each combination was estimated using Generalizability theory (35). For the Angoff data, analysis was based on the percentage of the total marks required for each question provided by the judges. The Angoff standards were provided by a different panel of judges for each station, i.e. the design of the G-analysis is a one facet random effect J: S design, where Judges (J) are nested within Stations (S). Variance components associated with the station, and the judge within station effects were obtained in an analysis of variance. Standards were obtained for the set of stations in the particular test (OSCE) at hand, so stations were considered a fixed facet and, as a consequence, the station variance does not enter the RMSE of the standard.
### TABLE 2. OSCE Checklist scores (mean and SD) and Global OSCE scores (mean and SD) per station of 119 dental students with pass/fail standards per station and per cluster (mean) of the Angoff I, the Angoff II, and the Borderline Regression (BR) method. And also the pass rates (% of students that passed) of these methods in 3 Compensatory models: Non Compensatory (NC), Partial Compensatory (PC) and Total Compensatory (TC)

<table>
<thead>
<tr>
<th>Clusters and stations</th>
<th>Checklist Scores Mean (SD) (%)</th>
<th>Global rating Mean (SD) (%)</th>
<th>Pass/fail standard</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clusters scores</td>
<td>Global rating</td>
<td>Angoff I (%)</td>
</tr>
<tr>
<td>Cluster 1 Professionalism</td>
<td>3. Cross infection control</td>
<td>72.1 (20)</td>
<td>61.2 (26)</td>
</tr>
<tr>
<td></td>
<td>4. Electronic Patient Record</td>
<td>69.9 (25)</td>
<td>60.4 (24)</td>
</tr>
<tr>
<td></td>
<td>12. Taking X-ray</td>
<td>79.9 (16)</td>
<td>76.4 (16)</td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster 2 Communication</td>
<td>5. History taking</td>
<td>70.8 (18)</td>
<td>60.8 (20)</td>
</tr>
<tr>
<td></td>
<td>8. Educating patient</td>
<td>65.3 (20)</td>
<td>56.2 (22)</td>
</tr>
<tr>
<td></td>
<td>10. Reflection feelings</td>
<td>71.3 (23)</td>
<td>63.6 (18)</td>
</tr>
<tr>
<td></td>
<td>13. Writing prescription</td>
<td>69.3 (20)</td>
<td>61.2 (24)</td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster 3 Diagnostics</td>
<td>2. Health risk ASA score</td>
<td>68.9 (25)</td>
<td>56.6 (22)</td>
</tr>
<tr>
<td></td>
<td>6. Periodontal index</td>
<td>72.9 (18)</td>
<td>61.4 (24)</td>
</tr>
<tr>
<td></td>
<td>9. Diagnostics caries</td>
<td>66.1 (15)</td>
<td>54.2 (20)</td>
</tr>
<tr>
<td></td>
<td>14. Tracing radiograph</td>
<td>54.6 (16)</td>
<td>42.6 (18)</td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster 4 Health maintenance</td>
<td>1. Resuscitation</td>
<td>76.6 (13)</td>
<td>54.8 (18)</td>
</tr>
<tr>
<td></td>
<td>7. Measuring pockets</td>
<td>75.5 (13)</td>
<td>62.8 (16)</td>
</tr>
<tr>
<td></td>
<td>11. Restoration</td>
<td>72.1 (19)</td>
<td>63.0 (22)</td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean total OSCE (TC model)</td>
<td>70.4 (8)</td>
<td>60.0 (8)</td>
<td>64.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pass rates %</th>
<th>Angoff I</th>
<th>Angoff II</th>
<th>BR</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC</td>
<td>0.8</td>
<td>1.7</td>
<td>7.6</td>
</tr>
<tr>
<td>PC</td>
<td>30.3</td>
<td>34.5</td>
<td>61.3</td>
</tr>
<tr>
<td>TC</td>
<td>86.6</td>
<td>86.6</td>
<td>97.5</td>
</tr>
</tbody>
</table>

OSCE = Objective Structured Clinical Examination

The pass rates for the clinically ‘incompetent’ and the clinically ‘competent’ students were calculated for all combinations of the three standard setting methods and the three compensatory models. The Weighted Loss was calculated for each combination in order to investigate their credibility.
Results
Table 2 shows the result of the OSCE: the mean (SD) checklist scores and mean (SD) global scores and the different pass/fail standards of the stations and the clusters and the total mean. No difference was found between the pass/fail standards across stations of the Angoff I and the Angoff II. The BR pass/fail standards were found to be lower than the Angoff methods (n=14, p<0.01). Additionally in the last rows of Table 2 the pass rates are shown of the 3 standard setting methods in the 3 compensatory models. Comparing the different compensatory models, a model without compensation (NC), will provide pass rates of 0.8%, 1.7% and 7.6%. The highest percentage of students will pass when the total compensatory model was used.

The Reliability of the pass/fail standards within two compensatory models.
Since the NC model gave extremely low pass rates, the reliability in this model was not estimated. Table 3 shows the RMSE of the pass/fail standards in the four clusters and the pass/fail standards with their 95% confidence intervals for the Angoff I and Angoff II and BR methods in the PC and the TC compensatory model. The RMSE of the standards in the Angoff methods is based on the number of judgements provided by the judges. This included 26 station-judgements provided by the judges for the clusters 1, 3, and 4 and 32 station-judgements for cluster 2. The total OSCE, (when compensating for all stations in the TC model), included 110 station-judgements. The RMSE of the standards in the BR method was based on the 119 students in the OSCE. The 95% confidence intervals are about a factor 3 smaller for the BR standard setting method in each cluster of competence compared to the Angoff methods.

Analyses showed that for the Angoff methods the RMSE decreases (thus the reliability increases) with increasing number of judgements of judges. This is visualized for the Angoff II method in Figure 2. For the BR method Figure 3 shows that with an increasing number of students in the OSCE the RMSE decreases, (thus the reliability increases).

The Credibility of the pass/fail standards within two compensatory models.
Since the NC model gave extremely low pass/fail standards (see Table 2) the credibility data from this model are not presented. The clinical rating of the sample in the dental training practice

<table>
<thead>
<tr>
<th>RMSE (%)</th>
<th>Pass/fail standard with 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster (PC)</td>
<td>Angoff I</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2.08</td>
</tr>
<tr>
<td>3</td>
<td>3.74</td>
</tr>
<tr>
<td>4</td>
<td>2.76</td>
</tr>
<tr>
<td>All stations (TC)</td>
<td>1.35</td>
</tr>
</tbody>
</table>
Fig. 2. Root mean squared error (RMSE) % of the pass/fail standards in the Angoff II standard setting method (with reality check) per cluster of competence: in Partial Compensatory model (PC) and for all stations in the Total Compensatory model (TC) as function of the number of judges (Nstations x Njudges/station).

Fig. 3. Root mean squared error (RMSE) % of the pass/fail standards in the Borderline Regression standard setting method per cluster of competence, in Partial Compensatory model (PC) and for all stations Total Compensatory model (TC) as function of the number of students. The RMSE for the clusters 1 and 4 are exactly the same for numbers of students, therefore the two lines overlap.
Standard setting in OSCE

resulted in a group of 14 clinically “incompetent” students and of 75 clinically “competent” students. Table 4 shows the pass rates of the incompetent and competent students using the different standard setting methods in the PC and the TC model. The TC model resulted in all 3 standard setting methods for the competent students in a pass rate >90 %, which is in contrast with the PC model where the pass rates were < 34% for both Angoff methods and 65% for the BR method. The TC model resulted for the incompetent students in pass rates of 50% for the Angoff methods and 92 % for the BR method. This is in contrast with the PC model, where for the incompetent students pass rates of 14% - 28% were obtained. In addition, no differences were found between the Angoff I and II in terms of pass rates.

To elucidate which method is the best to use in order to prevent incompetent students to pass (to protect the patients) and also to prevent the decision that competent students fail, a weighted loss was calculated on the basis of data of Table 4. The results are presented in Table 5. The formula presented in the Material and Methods section defines the weighted Loss in terms of the false positive rate and the true positive rate. By assigning a larger weight ($V=1, 10, ..., 100$), the decision of passing an incompetent student has a larger impact on the weighted Loss. The higher the value of the Loss, the greater the impact of this incorrect decision will be for the quality of future dental treatments. A

### Table 4. Pass rate % of students (N=89) who were rated by their clinical instructors as Incompetent and Competent students for the Angoff I and Angoff II and the Borderline Regression standard setting method in 2 Compensatory models: Partial Compensatory (PC), and Total Compensatory (TC)

<table>
<thead>
<tr>
<th>Standard setting method and model</th>
<th>Pass rate %</th>
<th>Competent students</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Incompetent students</td>
<td>n=14 (16%)</td>
</tr>
<tr>
<td>Angoff I PC</td>
<td>14.3</td>
<td>28.0</td>
</tr>
<tr>
<td>TC</td>
<td>50.0</td>
<td>90.7</td>
</tr>
<tr>
<td>Angoff II PC</td>
<td>14.3</td>
<td>33.3</td>
</tr>
<tr>
<td>TC</td>
<td>50.0</td>
<td>90.7</td>
</tr>
<tr>
<td>BR PC</td>
<td>28.6</td>
<td>65.3</td>
</tr>
<tr>
<td>TC</td>
<td>92.9</td>
<td>98.7</td>
</tr>
</tbody>
</table>

### Table 5. The weighted loss, the weighted sum of sensitivity and specificity, of the Angoff I and Angoff II and BR standard setting methods in the Partial (PC) and Total Compensatory (TC) model. The decision of passing of an incompetent student is $x$-times (weight) worse than failing of a competent student.

<table>
<thead>
<tr>
<th>Weight</th>
<th>FP= false positive and TP= true positive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Method</td>
<td>Model</td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>Angoff I</td>
<td>PC</td>
</tr>
<tr>
<td></td>
<td>TC</td>
</tr>
<tr>
<td>Angoff II</td>
<td>PC</td>
</tr>
<tr>
<td></td>
<td>TC</td>
</tr>
<tr>
<td>BR</td>
<td>PC</td>
</tr>
<tr>
<td></td>
<td>TC</td>
</tr>
</tbody>
</table>
weight $V=1$ corresponds to the loss associated with a false positive being considered equal to the loss associated with a false negative. For this situation (Table 5, fifth column) the loss for the PC model was found to be larger than the loss for the TC model. However, if the loss associated with a false positive was considered to be 10 or more times the loss of a false negative, the loss for the PC model was found to be lower than the loss of the TC model (Table 5, columns 6-10).

**Discussion**

This study tested 3 standard setting methods (Angoff I, Angoff II with a reality check and the Borderline Regression (BR) standard setting method), in order to establish an optimal choice for setting pass/fail scores for dental students participating in a dental OSCE.

In terms of reliability, the results of the present study show that the BR standard setting method provides a more reliable pass/fail standard than the Angoff methods in all compensatory models. In the BR method the RMSE and the related confidence intervals of the pass/fail standards are to some extend dependent on the number of students. In the present study 119 students participated in the OSCE. Where in comparable studies of Wood et al. (18) and Kramer et al. (23), a number of 58 and 89 students participated respectively, all 3 studies gave comparable RMSE's, and comparable 95% Confidence Intervals. It may be suggested that only 50 students would suffice to obtain a reliable pass/fail standard with the BR method.

Next to the reliability, this study also compared the credibility of the pass/fail standards for the combinations of the three standard setting methods and three compensatory models. The credibility was indicated by the weighted loss. Less loss means that there are less incorrect decisions made about passing and failing a student. In the PC model less loss was found in weights higher than 10. Since the BR method was more reliable, it seems justified to focus on the TC and PC models when using the BR. Since the TC model showed more loss than the PC model when applying the BR method, it can be concluded that the TC model gives to many incorrect decisions and for the BR method the PC model is the most credible model and can be used with better confidence.

Credibility of an OSCE is also determined by the leniency of the standard setting method that is used. Higher pass rates of a method implicates more leniency of the pass/fail standard applied. The pass rates of the BR method in the all compensatory models are higher than the pass rates of the Angoff methods (Table 2), and the reality check in the Angoff II method had no influence on the pass rates. As Cusimano also concluded (7) in his review about standard setting methods, there is considerable variability in the standards set by judges using different methods. In the present study the BR method is more lenient, giving higher pass rates, than the Angoff methods. These results are in accordance with the findings of Kane (21) and of Kramer et al. (23), They showed also that in a total compensatory model the Borderline Regression method was also more lenient than the Angoff method. However, in contrast to the study of Kramer et al., in our study the reality check in the Angoff method had no effect on the leniency.

In contrast to the BR method, an advantage of the Angoff I method is that the pass/fail standard is known for the students before the OSCE is administered. However, since judges in the Angoff I method tend to set this standard too high, a reality check is needed, by providing the judges with
the real scores of the students. And that is not possible before the OSCE is administered, which diminishes this advantage.

A drawback of the Angoff method is the time involved of the use of experts in panels, and therefore the costs. This was confirmed by the comments the judges made about the time consuming procedure. The judges needed about 2 hours for this procedure. Although the results of this study showed that more judgments of judges in the Angoff methods are resulting in higher reliabilities, the number of available judges is limited. Because the judgments in the BR method are made during the OSCE, this method is less expensive and may therefore be more acceptable.

When setting the standard, the purpose of the test must also be considered (7). The consequences of a false positive result after a periodontal course where in the next course of comprehensive care re-evaluation and re-education will be available is less damaging than the consequences of a false-positive result during a high-stakes final dental examination. The costs of remediation must be weighed in the balance, where in the other instance the cost to society of certifying an incompetent dentist must be weighed. In our study the OSCE tested a comprehensive care course for students finalizing the Bachelor phase and moving to the Master phase, where they will start with more complex dental treatments. The students must be competent in the basic clinical competencies, so that they can be trusted. In this study the results of the credibility and reliability of the Borderline regression method are giving evidence for this trust.

It can be questioned, to what extent the clinical rating that has been used as the criterion measure is a valid procedure. It is well known that this is a subjective assessment. Although it is far from perfect, this is the best criterion measure available and it is universally used as measure of clinical competence.

This study was based on only one sample of 119 students, which makes generalisations difficult. However, since this study confirms the results of the study of Kramer et al. the conclusion of the present study is that the use of the Borderline Regression method resulted in more reliable and more credible pass/fail standards than those of the Angoff methods. In addition the partial compensation within clusters of competence seems more credible than total compensation. Therefore the Borderline Regression method in a partial compensatory model provides defensible pass/fail standards and seems to be the optimal choice for this type of dental OSCE.
References