Study on the Interrater Reliability of an OSPE (Objective Structured Practical Examination) – Subject to the Evaluation Mode in the Phantom Course of Operative Dentistry

Abstract

Introduction: The aim of the study presented here was to evaluate the reliability of an OSPE end-of-semester exam in the phantom course for operative dentistry in Frankfurt am Main taking into consideration different modes of evaluation (examiner's checklist versus instructor's manual) and number of examiners (three versus four).
Methods: In an historic, monocentric, comparative study, two different methods of evaluation were examined in a real end-of-semester setting held in OSPE form (Group I: exclusive use of an examiner's checklist versus Group II: use of an examiner's checklist including an instructor's manual). For the analysis of interrater reliability, the generalisability theory was applied that contains a generalisation of the concept of internal consistency (Cronbach's alpha).
Results: The results show that the exclusive use of the examiner's checklist led to higher interrater reliability values than the in-depth instructor's manual used in addition to the list.
Conclusion: In summary it can be said that the examiner's checklists used in the present study, without the instructor's manual, resulted in the highest interrater reliability in combination with three evaluators within the context of the completed OSPE.

Keywords: OSCE, OSPE, checklist, evaluator, instructor's manual, feedback, dentistry

Laura Schmitt[1]
Andreas Möltner[2]
Stefan Rüttermann[3]
Susanne Gerhardt-Szép[3]

1 Goethe-University Frankfurt am Main, Carolinum Dental University Institute GmbH, Department of Orthodontics, Frankfurt/Main, Germany

2 University Heidelberg, Medical Faculty, Competence Centre for Examinations in Medicine/Baden-Württemberg, Heidelberg, Germany

3 Goethe-University Frankfurt am Main, Carolinum Dental University Institute GmbH, Department of Operative Dentistry, Frankfurt/Main, Germany

Introduction and Problem Definition

Performance checks constitute a central element of teaching; their evaluation is characterised primarily by the quality criteria of objectivity, reliability and validity [1], [2]. A GMA (Society for Medical Education) guideline [1] existing for this purpose and the basic standards of the WFME (World Federation for Medical Examination) [3] indicate the following criteria:

- the examinations must be justiciable
- the examination procedure is based upon learning goals and the learning effect on students
- the examination procedures applied and the guidelines for passing the exams must be made known.

In 2008, the Science Council recommended the creation of a functioning evaluation system on an international level for performance checks in universities. The task of the assessment tools applied was to analyse teaching performance clearly and dependably [http://www.wissenschaftsrat.de/download/archiv/8639-08.pdf, cited at 23.10.2015]. On the other hand, the current regulations on the licensing of dentists from 1955 contain no guidelines on the examinations held in the course of studies [http://www.gesetze-im-internet.de/z_pro/BJNR000370955.html, cited at 23.10.2015].
Because in the study of dentistry practical skills are reinforced, and thus also examined, we frequently deal with the implementation of competence-orientated methods of examination that can be characterised on the Miller pyramid by "shows how" or "acts" [4]. From this context, OSCE (Objective Structured Clinical Examination) and OSPE (Objective Structured Practical Examination) methods of examination are especially possible [4].
The OSCE method of examination was introduced in 1975 by Harden [5]. Initially conceived for examinations in

medicine, today the OSCE is also used for examinations in dentistry. In a 1998 study, Mangour and Brown [6] presented the development and implementation of OSCEs in dentistry for the first time. The terms OSCE and OSPE are usually applied as equivalents and thus with no differentiation. Both Natkin and Guild [7], as well as the AMEE (Association for Medical Education in Europe) Guide No. 81 Part I [8] describe OSPE (as a variation of OSCE) as a method of examination used to test practical skills and knowledge in a non-clinical environment. The authors Wani and Dalvi [9] also noted that the OSPE is an exam form where both the strengths and weaknesses of students' practical skill can be presented and reviewed. Students and examiners evaluate this exam form as positive and useful [1], [10], [11], [12], [13], [14]. In further studies, such as those of Smith et al. [15], Nayak et al. [16] and Abraham et al. [12], students described both OSCEs and OSPEs in comparison to written and oral examinations as fairer and less stressful exam forms, and preferred the OSPE to more "traditional" exam forms. A study by Schoonheim-Klein et al. [17] was also able to show that OSCEs, in a dental context in particular, promoted skills in the area of clinical competence and learning, as well as a more realistic self-assessment on the part of the students. In addition, the study by Nayak et al. [16] was able to show that through the OSPE, as well as the individual competencies of each student, the practical demonstration of facts and applied knowledge and learning behaviour could be positively influenced.

Reliability values between 0.11 and 0.97 were given for the OSCEs [18]. The strongly varying results can be explained primarily by the fact that the parameters under which an OSCE is held (number of stations, number of examiners, length of the exam, type of evaluation mode) could be seen to vary considerably.

Independently of the exam form, a differentiation is normally made in evaluation between the methods of "glance and grade" and evaluation based upon defined criteria. These methods were evaluated within the context of dental examination settings [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31]. The majority of the studies referred to above were not able to determine any significant differences between glance and grade and criteria-based methods. Furthermore, they did not take place in a real, but rather an artificial exam environment.

There are hardly any studies on OSPEs which, as already mentioned, represent in the strict sense a variation of the OSCE on the assessment of parameters referred to above. It has not been investigated, for instance, to what extent the number of examiners and the type of evaluation methods influence the result of an OSPE.

Against this background, the aims of this study were to evaluate the reliability of a real OSPE end-of-semester exam in the phantom course of operative dentistry in Frankfurt am Main, taking various evaluation modes and number of examiners into consideration.

Material and Methods

The phantom course of operative dentistry ran for a period of one semester (16 weeks). During this time, students had to complete practical work on a variety of simulation models (on extracted human and industrially manufactured artificial teeth). By means of previously defined treatment protocols, various treatment alternatives (for example fillings, laboratory restorations such as inlays, endodontic treatments, etc.) were practised step by step with the help of instructors. As soon as the predefined criteria were fulfilled, each step was ratified by the supervising instructor in a so-called certification booklet. The learning process was accompanied by formative feedback. At the end of the course, both an oral test of knowledge and a summative OSPE took place. The latter was carried out in the simulation unit of so-called "phantom patients". Two plastic models (upper and lower jaw) were mounted in a "phantom head" consisting of 14 plastic upper jaw teeth and 14 plastic lower jaw teeth. The OSPE consisted of two examination parts, the "filling" (A) and the "inlay" (B), carried out on two different plastic teeth of each respective model. These divided into six "sub-units" (1. "primary preparation"; 2. "under filling and secondary preparation"; 3. "filling"; 4. "inlay"; 5. "filling overall" and 6. "overall grade") which were each evaluated by the examiners (see Figure 1). These subunits accorded to the criteria based on which the attendance certificates for the course were issued by the instructors. The examiner's checklist, which contained the list of partial aspects (subunits) mentioned above, was tested out over four consecutive semesters (summer semester 2008 to winter semester 2009) in a regular examination scenario. During the test, the evaluation took place via inspection of the prescribed partial aspects, judged purely on the basis of the view of the examiners' general quality criteria. School grades were awarded from 1 to 5 (1=very good to 5=insufficient).

Each examiner evaluated each student in a real examination scenario (duration: 3 hrs.). This meant that the examiners assessed the students' work directly at the workplace (on a phantom patient) in a predetermined order during the examination. The students signalled to the examiners that they were ready to submit a subunit for evaluation. During the OSPE, the examiners exchanged no information on the grades they had awarded. After the examiners had independently completed their individual examiner's checklists, the evaluations were discussed in a joint meeting and it was determined which students should repeat the exam. This took place according to the Delphi principle [http://www.horx.com/zukunftsforschung/Docs/02-M-09-Delphi-Methode.pdf, cited at 23.10.2015].

Examination scenario of the study

The present study relates to a period of two semesters (summer semester 2010 = Group I, summer semester

A.			**Filling**
	I		Primary preparation
		1.	Proximal contact
		2.	Surface flat and smooth
		3.	Angle axial wall of the pulp
		4.	Breadth and depth
	II		Under filling and secondary preparation
		1.	UF smooth and
		2.	UF localisation
		3.	UF height
	III		Filling
		1.	Marginal seals
		2.	Contact points
		3.	Occlusal design
		4.	Smoothness
			Filling total
B			**Inlay**
	I	1.	Cavity outer edges
		2.	Cavity inner walls
		3.	Breadth and depth
		4.	Smoothness
		5.	Adjacent tooth
			Inlay total
			OVERALL GRADE OSPE

Figure 1: Examiner's checklist group I and group II with both tasks A (filling) and B (inlay). The abbreviation UF stands for under filling.

Table 1: Composition of the study population.

Group	Number of Students	Women	Men	Average Age (Years)
I	64	38	26	24,9
II	44	27	17	24,7

2012 = Group II). The composition of the study population is given in Table 1. The inclusion criteria were:

- students from the 6th semester
- participation in the phantom course for restorative dentistry
- examination skills present.

The exclusion criteria were defined as follows:

- students from other semesters
- course dropouts and course repeaters
- examination skills not met.

The difference in the respective group sizes (I versus II) resulted from the actual size of the semester which was subject to large variations and which was dependent upon the results of the preceding examination. A numerical adjustment of both groups was not feasible as all course participants, according to the study regulations, had to take the exam. The determination of the number of examiners was carried out prior to this study on application for ethical approval. The assignment of identical examiners for both groups was not practical for staffing reasons in the department.

In group I, an examiner's checklist was applied exclusively, as seen in Figure 1. In group II, the examiners used the identical examiner's checklist, but in combination with a detailed instructor's manual (see Figure 2). This contained clearly defined criteria for the evaluation the individual school grades.

In all, five examiners took part in the study (A-E), four women and one man. The examiners were all dentists in the Department for Operative Dentistry, had experience in teaching and in the evaluation of students' work in the phantom course. Table 2 shows their distribution according to number and sex. Examiner A had passed the final examination in dentistry in 1990, examiners B, C, D, and E in 2007, 2008, 2010 and 2011 respectively. They all had experience in conducting the phantom course of operative dentistry. In addition to the others, only A had experience in conducting courses in patient treatment.

The examiner's checklist originated from subject areas that were presented as standard in the current course, and in textbooks for restorative dentistry. These were also similar to the units (filling, inlay) and subunits defined as relevant for examination in operative dentistry raised in Baumann's study [32] on an interdisciplinary basis between four centres (the universities of Frankfurt,

Primary Preparation

1. **Contact proximal**
   1. All 3 surfaces probe tip dissolved
   2. 1 surface double probe tip dissolved
   3. 1 surface probe tip and 2 surfaces double probe tip dissolved
   4. All 3 surfaces double probe tip dissolved
   5. At least 1 surface more than double probe tip, or at least 1 surface not dissolved
2. **Base flat and smooth**
   1. Optimally flat and smooth
   2. +/- 5 degrees flat and smooth
   3. Flat or smooth, +/- 5 degrees flat and rough (smoothing with Arkansas was necessary)
   4. +/- 19 degrees and smooth
   5. > 10 degrees of flat
3. Angle pulpal axial wall
   1. Broken and 90 degrees/slightly conical
   2. Not broken and 90 degrees/slightly conical
   3. Not broken and undercut (5 degrees)
   4. Diverging by > 5 degrees
   5. Undercut/diverging > 5 degrees
4. Breadth and depth
   1. Optimally 10 cone sunk, breadth 1/3 of the buccal and oral cusp spacing
   2. Depth > than ¾ of the cusp can still be sunk, breadth 1/3 of the buccal and oral cusp spacing
   3. Breadth and depth do not deviate by more than 10%
   4. Breadth and depth do not deviate by more than 15%
   5. Filling not possible / P

Under filling and Secondary Preparation

1. **UF smooth and transition rounded**
   1. Optimally flat and smooth, transition broken
   2. < 5 degree deviation from flat, edges rounded
   3. Maximal 5 degree deviation from flat, edges rounded
   4. 10 degree deviation from smooth, edges rounded
   5. > 10 degree deviation from smooth sharp edges
2. **UF localisation**
   1. Enamel margins free, only on p.a. wall and box floor
   2. Enamel margins free, 90% p.a. wall and box floor covered by UF
   3. Enamel margins free, 80% of p.a. wall and box floor covered by UF
   4. Enamel margins partly free but more than only p.a. wall and box floor
   5. P.a. wall is exposed >20% and/or phosphate cement on enamel margins
3. **UF height**
   1. ¾ of the cusp can be sunk, optimally flat and smooth
   2. ¾ of the cusp can be sunk, optimally flat, slight unevenness
   3. ¾ of the cusp can be sunk, +/- 5 degrees flat and smooth
   4. ¾ of the cusp can be sunk, +/- 10 degrees flat and smooth
   5. < than ¾ of the cusp can be sunk

Filling

1. **Border seal**
   1. Filling tight, no crack
   2. Filling approximately tight, probe remains minimally hanging
   3. Filling slightly over/under contoured
   4. Crack <0.5mm
   5. Crack clearly palpable
2. **Contact points**
   1. optimal (matrix band with impression)
   2. Minimally too strong/weak (matrix band no impression)
   3. Dental floss holds
   4. Double dental floss holds
   5. Crack visible (double dental floss does not hold) proximal contact blocked
3. **Occlusal design**
   1. Anatomical design
   2. Anatomically similar design
   3. Contouring at least 1 main fissure & 2 side bars
   4. Contouring at least 1 main fissure
   5. No contouring recognisable
4. **Smoothness**
   1. Polishing complete, high polish recognisable
   2. Polishing complete, polish recognisable
   3. Polishing complete
   4. Filling not polished
   5. Pores, inclusions, no contouring with rotating instruments

Inlay

1. **Cavity outer edges**
   1. Dimension and prox. optimal dissolving (double probe tip)
   2. Dimension and prox. dissolving near optimal (probe tip size)
   3. Preparation too broad/too prox. dissolving at least probe tip size
   4. Defined preparation margin line, prox. dissolving at least probe tip size
   5. Impression and inlay not possible
2. **Cavity inner walls**
   1. Conical and rounded
   2. Defects in conical or rounded
   3. Walls conical but not rounded
   4. Cavity rounded, walls only approximately conical (inlay possible in laboratory)
   5. Inlay not possible
3. **Breadth and depth**
   1. Optimal depth at least 1.5mm, breadth 1/3 of the buccal/oral cusp spacing
   2. Approximately optimal (depth between 1.5-2mm, breadth 1/3 of the buccal/oral cusp spacing
   3. Depth between 1.5-2mm 10% deviation from 1/3 of the buccal/oral cusp spacing
   4. Depth max. 2.5mm, breadth 10% deviation from 1/3 of the buccal/oral cusp spacing
   5. Depth less than 1.5mm, isthmus too narrow/too broad (>10% deviation from 1/3 of the buccal/oral cusp spacing)
4. **Smoothness**
   1. Optimally flat and smooth
   2. Optimally flat and approx. smooth
   3. Occasional unevenness
   4. Scratches
   5. No smoothing occurs
5. **Adjacent teeth**
   0 not prepared
   1 grade deduction one side prepared in enamel
   2 grades deduction both prepared in enamel
   Failed one prepared in dentine
   Failed both prepared in dentine

Overall grade

Figure 2: Instructor's manual for group II with the evaluation criteria of both tasks A and B. The abbreviation p.a. signifies pulpal axial wall. The abbreviation UF stands for under filling, prox. = proximal.

Table 2: Data for the examiners A to E, who evaluated group 1 in SS 2010 (A, B, C) and group 2 in SS 2012 (A, B, D, E) (f = female, m = male).

Group	Examiners	Number of examiners	Gender of the examiners (f)	Gender of the examiners (m)
I	A, B, C	3	2	1
II	A, B, D, E	4	3	1

Freiburg, Leipzig and Munich). From the manual attached to group II, examiners were able to learn which evaluation criteria had to be fulfilled in order for a particular grade to be awarded.

Train-the-Teacher

In each semester, a 45 minute "train-the-teacher course" was held. In this course, examiners were prepared through practical exercises and theoretical instructions on situations in the OSPE and the use of the examiner's checklist and the instructor's manual. Thus in advance a relatively high measure of standardisation between the examiners could be achieved.

Statistics and Application for Ethical Approval

The results were evaluated according to the generalisability theory (G theory) with the statistic programmes SAS 9.2 (SAS Institute Inc., Cary, USA, PROC MIXED) and R (Version 2.15, Package lme4). The variance of the grades obtained is attributed to the influencing factors (in the terminology of the G theory "facets") "students" and "examiners", as well as to a measurement error component (see Figure 3). From the variance proportions of the facet "examiner" and error variance relative to the facet "student", the measurement reliability of the evaluations can be estimated. The generalisability coefficient represents an analogue to internal consistency (Cronbach's alpha). In contrast to its usual application to various tasks, it is used here for several examiners. The G theory allows assessment of measurement reliability with the adoption of a different number of examiners to that in the actual investigation. In this way, both studies in which a varying number of examiners were involved can be made compatible (in analogy to the Spearman-Brown formula with which a standardisation of reliability for a certain number of tasks is possible).
Similarly, the individual examiners (A-E) were evaluated amongst themselves with regard to the parameter "overall grade OSPE". A sub-group analysis taking in all parameters of examiners A and B completed the statistical analysis.
An application for ethical approval for the monocentric comparative study was given the approval number 135/35 by the Ethic Commission of the Department of Medicine of the Goethe University

Results

Table 3 shows the results of the determination of reliability from group I using the examiner's checklist without the instructor's manual. In this group, only in the case of three examiners were Cronbach's alpha values under 0.6 determined for the two criteria "interior wall of the cavity" and "breadth/depth".
In all other subunits, the required value of 0.6 or larger than 0.6 for sufficient reliability could be attained. The subunit "adjacent tooth" achieved the value 1.0; this can be regarded as an ideal reliability value. Furthermore, table 3 shows the results of the determination of reliability from group II (using the examiner's checklist and the instructor's manual). In order to enable a comparison of the generalisability coefficients in both studies, these were each converted for numbers of both three and four examiners. Thus with the aid of the Spearman-Brown formula, for study I the reliability values for four examiners were determined from those for three examiners, and vice versa for group II.
In group II the results for 4 examiners showed a high variance in the calculated Cronbach's alpha values. For the first subunit "primary preparation" and the accompanying criteria ("proximal contact point" to "breadth/depth), Cronbach's alpha values under 0.6 were calculated. The same was the case for the subunit "filling" and the accompanying criteria "contact points", "occlusal design" and "smoothness", for "inlay total" and accompanying criteria such as "cavity outer edge", "cavity inner walls", "breadth/depth", "smoothness" and "adjacent tooth". The remaining subunits and criteria were able to achieve the required value for sufficient reliability of 0.6.
When comparing individual examiners regarding the parameter "overall grade OSPE", for the summer semester 2010, correlation coefficients of 0.58 (A versus C), 0.64 (A versus B) and 0.68 (C versus B) were calculated. In the summer semester 2012, the corresponding values were lower (A versus B: 0.33; A versus E: 0.35; A versus D: 0.34; E versus D: 0.52; B versus D: 0.37 and E versus B: 0.35). The results of the subgroup analysis (A versus B, used in both study groups) can be seen in table 3.

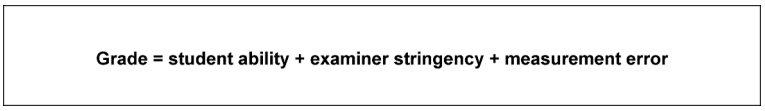

Figure 3: The facets of the variance analysis conducted in the study.

Table 3: Results of group I and group II. The corresponding reliability values (Cronbach's alpha) are given for three and four examiners. In the column "A vs. B", the results of the subgroup analysis are presented. Identification with * means that differing from the real examination scenario, a conversion into another number of examiners (abbreviations: CL = cavity lining, vs. = versus).

				Group I				Group II		
				3	4*	A vs. B		3*	4	A vs. B
A			**Filling**							
	I		Primary Preparation	0.867	0.897	0.667		0.373	0.443	0.111
		1.	Contact, proximal	0.642	0.705	0.450		0.311	0.376	-0.039
		2.	Floor flat and smooth	0.722	0.776	0.593		0.504	0.575	0.347
		3.	Pulp axial wall	0.696	0.753	0.433		0.513	0.584	0.381
		4.	Breadth and depth	0.84	0.875	0.676		0.382	0.451	0.086
	II		Base and secondary preparation	0.835	0.871	0.714		0.798	0.84	0.748
		1.	Base smooth and transition rounded	0.725	0.779	0.492		0.69	0.748	0.512
		2.	Baser localisation	0.689	0.747	0.714		0.56	0.629	0.474
		3.	Base height	0.691	0.749	0.547		0.753	0.803	0.610
	III		Filling	0.796	0.839	0.635		0.716	0.77	0.473
		1.	Marginal seal	0.663	0.724	0.507		0.666	0.726	0.346
		2.	Contact point	0.655	0.717	0.525		0.347	0.415	0.074
		3.	Occlusal configuration	0.718	0.772	0.543		0.371	0.44	0.179
		4.	smoothness	0.777	0.823	0.661		0.428	0.499	0.346
			Filling total	0.823	0.861	0.790		0.739	0.791	0.583
B			**Inlay**	0.72	0.774	0.407		0.406	0.476	0.098
	I	1.	Cavity, outer edges	0.678	0.738	0.356		0.102	0.132	0.179
		2.	Cavity, inner walls	0.548	0.617	0.305		0.311	0.376	-0.017
		3.	Breadth and depth	0.302	0.365	0.161		0.339	0.406	0.075
		4.	smoothness	0.661	0.723	0.431		0.278	0.339	0.082
		5.	Adjacent tooth	1	1	1.000		0	0	-0.168
			Inlay total	0.72	0.774	0.407		0.406	0.476	0.098
			OVERALL GRADE OSPE	**0.839**	**0.874**	**0.648**		**0.633**	**0.697**	**0.330**

Discussion

Limitations

One limitation of the present study lies in the type of trial design selected (historical comparison group), as the study was carried out not within one particular semester with a particular student population, but rather in two successive semesters with different participants. Because of two different modes of assessment, a division of the summative examination within the semester was declared inadmissible by the faculty's ethics commission. The authors see one further limitation in the fact that the examiners from both investigated groups were not equal either in number or team composition. Only two examiners (A and B) evaluated similarly in both study groups. Furthermore, despite the preceding train-the-teacher events, a difference in teaching experience must be assumed. This variation could, however, not be homogenised for staff reasons (expiry of contracts). The elaborate statistical analysis takes account of this limitation and standardises the unequal number of examiners.

Modes of evaluation

Based on current scientific information, no clear conclusion can be drawn on the benefit of an examiner's checklist regarding the reliability of an examination. According to the latest research, there are only two studies which have dealt with the different modes of evaluation [19], [20], [26], [28], [29], [33]. In the present study, the best results could be determined regarding a high level of reliability by using the examiner's checklist without the additional use of an instructor's manual. A comparable result was achieved in a study by Bazan and Seale [34], where a similarly conceived examiner's checklist for exam evaluation led to a similar reliability value for the exam. An explanation for this might be that the degree of differentiation in the evaluation guidelines was possibly too detailed to be applied by the examiner during the practical examination, and that the train-the-teacher event was apparently not able to set comparable evaluation standards for the examiners. This problem became particularly apparent in the partial step "inlay adjacent tooth" in which the extensive manual with the defined sub-criteria led to a massive deterioration in the Cronbach's alpha values. This is also accords with the study by the authors Houpt and Cress [31], which found that the narrower the definition of the predetermined evaluation framework for a criterion was, the sooner discrepancies in measurement accuracy and examiner assessment occurred. A direct comparison of examiners A and B, who examined in both semesters, found that the use of the manual lowered the average correlation (0.68) recorded in summer semester

2010 to a value of 0.33. Despite this clarification, it is still necessary to establish why this partial step in particular caused such extreme deviations. Possibly the wording of the tooth structure definitions (enamel and dentine) resulted in confusion on the side of the examiners as the exam tasks were not carried out on natural teeth consisting of enamel and dentine, but rather on exam teeth made of plastic. Future studies should discuss the exact wording of the manual parameters in terms of content.

Examination setting

In contrast to the two studies already referred to, the examiners' evaluation in the present study took place in a real exam situation. As a potential future alternative regarding study design, it would be feasible to give the examiners more time for evaluation. This, however, would require a fundamental revision of the end-of-semester exam at the University of Frankfurt am Main under study here. Considering that three hours were allowed for the whole examination, and that the individual steps were checked simultaneously ad hoc by the examiners with an average of = 22 students, more time spent on the evaluation could only be realised with difficulty. The question arises of why, during the real OSPE examination scenario, so much effort is expended and why the individual steps cannot be evaluated jointly by all the examiners after the exam. The reason for this is that many individual steps during the exam are no longer assessable owing to the succeeding phase, as they are then no longer visible. For example, the "primary preparation" step succeeding "under filling lining/secondary preparation" is no longer assessable as the former is partially concealed after putting in an under filling. This is the same for all partial steps so that at the end of the examination stage "filling", only the final resulting step remains assessable.
This procedure stands in stark contrast to all previously published OSPE examinations where in general the individual steps were both visible and assessable, even after the examination. Compared to the studies made by Goepferd and Kerber [26], Vann et al. [28] and Scheutzel [33] there is a clear difference, as in the examinations investigated there, the similarly complex revaluation form was able to be used under more favourable time conditions. This might explain the different results between the investigation carried out here and the studies previously referred to.

Train-the-Teacher

OSCE-based examinations show some disadvantages by way of analogy to the advantages already referred to above. According to Miller [4], [35], experience has shown that the OSCE is particularly training intensive and time consuming, and according to Nayak et al. [16], it requires intensive planning and team work. As a rule, the appointed examiners require intensive and systematic training in order to be able to fulfil the requirements of reliability and validity for an OSCE exam [35]. As a result, the OSCE is time consuming and cost intensive in comparison to other exam types such as multiple choice or oral exams [8], [35], [36]. In the context of the present study, a time-consuming preparation of the examiners in a train-the-teacher event was also carried out. As a result, resources of personnel and space, as well as financial resources in the clinical and organisational workflow within the department for restorative dentistry, would have to be found. The duration of a lecture unit (45 mins.) was realistic for this purpose and could be observed by all the examiners. However, the question arises as to how long preparation should effectively be in order to be able to homogenise different experiences in mixed teams in advance. In the summer semester of 2010, the three examiners amongst themselves showed an average correlation of between 0.58 and 0.68. In the summer semester of 2012, in the case of four examiners the identically long train-the-teacher events resulted in correlation values of 0.33 and 0.52. It can be assumed here that in the case of the application of the manual, the train-the-teacher event was not effectively utilised.

Examiners

On the basis of current data, examiners play an important role in the assessment of reliability. Until now, however, there have been no scientific studies known to us that have made any assessment of how high the minimum number of examiners for a OSPE should be. In this study, it was possible to attain sufficient reliability with three examiners in combination with checklists. According to the results of this investigation, the reliability value can be increased by a higher number of examiners. This increase in reliability values, however, is low in comparison to the number of examiners. In addition, a further increase in the number of examiners would result in greater complexity and expense with regard to organisation and financial costs.
In this context, it has to be mentioned critically that no general recommendation can be made for other sites based upon the data available with regard to the number of examiners, as the possibility of having three to four examiners with long experience available for an OSPE examination is neither representative of normal circumstances nor feasible. The author groups Nikendei and Jünger [37] and Norcini et al. [38] came to a similar result. In their study, Natkin and Guild [39] were able to show a significant increase in reliability through a systematic preparation of the evaluators. Similar results were presented by Dhuru [25], in whose study examiners with many years of professional experience and using evaluation sheets achieved the most reliable examination results. In the present study, this can be confirmed only with the use of the checklist, as when the manual was used, the two examiners with the most years' experience demonstrated only weak correlations. As shown in this investigation, the checklist appears to be capable of further increasing reliability, or of compensating for a lack of examining experience on the part of the evaluators. In

Houpt and Kress's [31] investigation, by contrast, reliability could not be increased for all evaluation criteria. Thus the authors believe that the train-the-teacher events on their own are not able to increase interrater reliability significantly. Training events of this type had the greatest effect with "non-expert" examiners, but relatively little influence with experienced evaluators [31]. Our study was able to confirm this.

Exam tasks

The number of examination tasks defined in this study, frequently equated with the term "stations" in the literature, should be looked at critically. In the present case only two separate tasks were involved (A. filling and B. inlay), but a total of 22 evaluations were obtained by the evaluators per student in and during the exam. Ultimately we are dealing with the definition of the term "station" in connection with the OSPE which based upon the evidence cannot be deduced from the literature. It must be noted critically that a value of 0.6 for Cronbach's alpha only has a "sufficient" character. It must therefore also be asked just how valid an examination can then be, and whether it is suitable as a summative examination. According current scientific knowledge, it is our opinion that against this background, variant II cannot be recommend for high stakes examinations.

Conclusion

The following conclusions may be drawn from this study regarding the question of how an OSPE in dental teaching in a phantom course for operative dentistry can best be reliably designed:

- an examiner's checklist without an instructor's manual resulted in higher interrater reliability in the context of the OSPEs carried out
- the evaluation of students' exam performance in the context of the OSPE should if possible be undertaken by at least three examiners.

Acknowledgements

The authors would like to thank the students of the 6th semester in the section for operative dentistry and the dental course assistants who also contributed to the evaluation of the OSPE.

Competing interests

The authors declare that they have no competing interests.

References

1. Gesellschaft für Medizinische Ausbildung, Kompetenzzentrum Prüfungen Baden-Württemberg, Fischer MR. Leitlinie für Fakultätsinterne Leistungsnachweise während des Medizinstudiums: Ein Positionspapier des GMA-Ausschusses Prüfungen und des Kompetenzzentrums Prüfungen Baden-Württemberg. GMS Z Med Ausbild. 2008;25(1):Doc74. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2008-25/zma000558.shtml
2. Taylor CL, Grey NJ, Satterthwaite JD. A comparison of grades awarded by peer assessment, faculty and a digital scanning device in a pre-clinical operative skills course. Eur J Dent Educ. 2013;17(1):16-21. DOI: 10.1111/j.1600-0579.2012.00752.x
3. World Federation for Medical Education. Basic Medical Education The 2012 Report. Copenhagen: WFME Office; 2012.
4. Miller GE. The assessment of clinical skills/competence/performance. Acad Med.1990;65:S63-67. DOI: 10.1097/00001888-199009000-00045
5. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J. 1975;1:447-451. DOI: 10.1136/bmj.1.5955.447
6. Manogue M, Brown G. Developing and implementing an OSCE in dentistry. Eur J Dent Educ.1998;2(2):51-57. DOI: 10.1111/j.1600-0579.1998.tb00039.x
7. Natkin E, Guild RE. Evaluation of preclinical laboratory performance: a systematic study. J Dent Educ.1967;31(2):152-161.
8. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. Med Teach. 2013;35(9):e1437-1446. DOI: 10.3109/0142159X.2013.818634
9. Wani P, Dalvi V. Objective Structured Practical Examination vs Traditional Clinical Examination in Human Physiology: Students perception. Int J Med Sci Public Health. 2013;2(3):522–547. DOI: 10.5455/ijmsph.2013.080320133
10. Schoonheim-Klein M, Muijtjens A, Muijtens A, Habets L, Manogue M, van der Vleuten C, Hoogstraten J, Van der Velden U. On the reliability of a dental OSCE, using SEM: effect of different days. Eur J Dent Educ. 2008;12(3):131–137. DOI: 10.1111/j.1600-0579.2008.00507.x
11. Hofer M, Jansen M, Soboll S. Potential improvements in medical education as retrospectively evaluated by candidates for specialist examinations. Dtsch Med Wochenschr. 2006;131(8):373–378. DOI: 10.1055/s-2006-932527
12. Abraham RR, Raghavendra R, Surekha K, Asha K. A trial of the objective structured practical examination in physiology at Melaka Manipal Medical College. India. Adv Physiol Educ. 2009;33(1):21–23. DOI: 10.1152/advan.90108.2008
13. Adome RO, Kitutu F. Creating an OSCE/OSPE in a resource-limited setting. Med Educ. 2008;42(5):525–526. DOI: 10.1111/j.1365-2923.2008.03045.x
14. Davenport ES, Davis JE, Cushing AM, Holsgrove GJ. An innovation in the assessment of future dentists. Br Dent J. 1998;184(4):192–195.
15. Smith LJ, Price DA, Houston IB. Objective structured clinical examination compared with other forms of student assessment. Arch Dis Child. 1984;59:1173-1176. DOI: 10.1136/adc.59.12.1173
16. Nayak V, Bairy KL, Adiga S, Shenoy S, Magazine BC, Amberkar M, Kumari M. OSPE in Pharmacology: Comparison with the conventional Method and Students' Perspective Towards. Br Biomed Bull. 2014;2(1):218-222.

17. Schoonheim-Klein ME, Habets LL, Aartman IH, van der Vleuten CP, Hoogstraten J, van der Velden U. Implementing an Objective Structured Clinical Examination (OSCE) in dental education: effects on students' learning strategies. Eur J Dent Educ. 2006;10(4):226-235. DOI: 10.1111/j.1600-0579.2006.00421.x

18. Chenot JF, Ehrhardt M. Objective structured clinical examination (OSCE) in der medizinischen Ausbildung: Eine Alternative zur Klausur. Z Allg Med. 2003;79(2):437-442.

19. Sharaf AA, AbdelAziz AM, El Meligy OA. Intra- and inter-examiner variability in evaluating preclinical pediatric dentistry operative procedures. J Dent Educ. 2007;71(4):540-544.

20. Kellersmann CT. Zur Reliabilität der Beurteilung vorklinischer Phantomarbeiten bei Einsatz eines strukturierten Bewertungsbogens. Inaugural-Dissertation. Münster: Westfälischer Wilhelms-Universität Münster; 2007.

21. Lilley JD, ten Bruggen Cate HJ, Holloway PJ, Holt JK, Start KB. Reliability of practical tests in operative dentistry. Br Dent J. 1968;125(5):194-197.

22. Fuller JL. The effects of training and criterion models on interjudge reliability. J Dent Educ. 1972;36(4):19-22.

23. Hinkelman KW, Long NK. Method for decreasing subjective evaluation in preclinical restorative dentistry. J Dent Educ. 1973;37(9):13-18.

24. Gaines WG, Bruggers H, Rasmussen RH. Reliability of ratings in preclinical fixed prosthodontics: effect of objective scaling. J Dent Educ. 1974;38(12):672-675.

25. Dhuru VB, Rypel TS, Johnston WM. Criterion-oriented grading system for preclinical operative dentistry laboratory course. J Dent Educ.1978;42(9):528-531.

26. Goepferd SJ, Kerber PE. A comparison of two methods for evaluating primary class II cavity preparations. J Dent Educ. 1980;44(9):537-542.

27. Feil PH. An analysis of the reliability of a laboratory evaluation system. J Dent Educ. 1982;46(8):489-494.

28. Vann WF, Machen JB, Hounshell PB. Effects of criteria and checklists on reliability in preclinical evaluation. J Dent Educ. 1983;47(10):671-675.

29. Bedi R, Lo E, King NM, Chan T. The effect of pictorial criteria upon the reliability of assessments of cavity preparations. J Dent. 1987;15(5):222-224. DOI: 10.1016/0300-5712(87)90116-3

30. Jenkins SM, Dummer PM, Gilmour AS, Edmunds DH, Hicks R, Ash P. Evaluating undergraduate preclinical operative skill; use of a glance and grade marking system. J Dent. 1998;26(6):679-684. DOI: 10.1016/S0300-5712(97)00033-X

31. Houpt MI, Kress G. Accuracy of measurement of clinical performance in dentistry. J Dent Educ. 1973;37(7):34-46.

32. Baumann MP. Evaluation von Bewertungskriterien für praktische Studentenarbeiten im Vergleich zur Bewertung per Augenschein. Inaugural-Dissertation. München: Medizinischen Fakultät der Ludwig-Maximilians-Universität München; 2015.

33. Scheutzel P. Einfluss des Bewertungssystems auf Objektivität und Reliabilität der Benotung zahnmedizinischer Studentenarbeiten am Phantompatienten. GMS Z Med Ausbild. 2007;24(1):Doc67. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2007-24/zma000361.shtml

34. Bazan MT, Seale NS. A technique for immediate evaluation of preclinical exercises. J Dent Educ. 1982;46(12):726-728.

35. Barman A. Critiques on the Objective Structured Clinical Examination. Ann Acad Med Singapore. 2005;34(8):478-482.

36. Boursicot K, Ware J, Hazlett C. Objective Structured Clinical Examination Objective Structured Practical Examination. Med Educ. 1979;31:41-54.

37. Nikendei C, Jünger J. OSCE-praktische Tipps zur Implementierung einer klinisch-praktischen Prüfung. GMS Z Med Ausbild. 2006;23(3):Doc47. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2006-23/zma000266.shtml

38. Norcini JJ, Maihoff NA, Day SC, Benson JA. Trends in medical knowledge as assessed by the certifying examination in internal medicine. JAMA. 1989;262(17):2402–2404. DOI: 10.1001/jama.1989.03430170064029

39. Natkin E, Guild RE. Evaluation of preclinical laboratory performance: a systematic study. J Dent Educ. 1967;31(2):152-161.

Corresponding author:
PD Dr. med. dent. Susanne Gerhardt-Szép, MME
Goethe-University Frankfurt am Main, Carolinum Dental University Institute GmbH, Department of Operative Dentistry, D-60596 Frankfurt/Main, Germany, Phone: +49 (0)69/6301-7505, Fax: +49 (0)69/6301-3841
s.szep@em.uni-freiburg.de

Please cite as
Schmitt L, Möltner A, Rüttermann S, Gerhardt-Szép S. Study on the Interrater Reliability of an OSPE (Objective Structured Practical Examination) – Subject to the Evaluation Mode in the Phantom Course of Operative Dentistry. GMS J Med Educ. 2016;33(4):Doc61. DOI: 10.3205/zma001060, URN: urn:nbn:de:0183-zma0010608

This article is freely available from
http://www.egms.de/en/journals/zma/2016-33/zma001060.shtml

Received: *2015-10-23*
Revised: *2016-04-01*
Accepted: *2016-06-03*
Published: *2016-08-15*

This is the translated (German) version.
The original (English) version starts at p. 1.

Studie zur Interrater-Reliabilität einer OSPE (Objective Structured Practical Examination) in Abhängigkeit vom Bewertungsmodus im Phantomkurs der Zahnerhaltungskunde

Zusammenfassung

Einleitung: Ziel der vorliegenden Studie war es, die Reliabilität einer OSPE-Semesterabschlussprüfung im Phantomkurs der Zahnerhaltungskunde in Frankfurt am Main unter Berücksichtigung unterschiedlicher Bewertungsmodi (Prüfer-Checkliste versus Dozentenmanual) und PrüferInnenanzahl (drei versus vier) zu evaluieren.
Methoden: Im Rahmen einer historischen monozentrischen Vergleichsstudie wurden zwei verschiedene Bewertungsmodi (Gruppe I: Verwendung ausschließlich einer Prüfer-Checkliste versus Gruppe II: Verwendung einer Prüfer-Checkliste inklusive eines Dozentenmanuals) im Rahmen einer realen Semesterabschlussprüfung, die in OSPE-Form abgehalten wurde, evaluiert. Zur Analyse der Interrater-Reliabilität wurde die Generalisierbarkeitstheorie verwendet, die eine Verallgemeinerung des Konzepts der internen Konsistenz (Cronbachs alpha) beinhaltet.
Ergebnisse: Die Ergebnisse zeigen, dass die alleinige Verwendung der Prüfer-Checkliste zu höheren Interrater-Reliabilitätswerten führte als das zusätzlich zu der Liste verwendete ausführliche Dozentenmanual.
Schlussfolgerung: Zusammenfassend kann festgehalten werden, dass die in der vorliegenden Studie verwendete Prüfer-Checkliste ohne Dozentenmanual im Rahmen der durchgeführten OSPE die höchste Interrater-Reliabilität ergab in Kombination mit der Anzahl von drei BewerterInnen.

Schlüsselwörter: OSCE, OSPE, Checkliste, Bewerter, Dozentenmanual, Feedback, Zahnmedizin

Laura Schmitt[1]
Andreas Möltner[2]
Stefan Rüttermann[3]
Susanne Gerhardt-Szép[3]

1 Goethe-Universität Frankfurt am Main, Carolinum Zahnärztliches Universitäts-Institut gGmbH, Poliklinik für Kieferorthopädie, Frankfurt/Main, Deutschland

2 Universität Heidelberg, Medizinische Fakultät, Kompetenzzentrum für Prüfungen in der Medizin/Baden-Württemberg, Heidelberg, Deutschland

3 Goethe-Universität Frankfurt am Main, Carolinum Zahnärztliches Universitäts-Institut gGmbH, Poliklinik für Zahnerhaltungskunde, Frankfurt/Main, Deutschland

Einleitung und Problemstellung

Leistungskontrollen bilden einen zentralen Bestandteil der Lehre; deren Evaluation wird in erster Linie durch die Gütekriterien Objektivität, Reliabilität und Validität charakterisiert [1], [2]. Eine hierzu existierende Leitlinie der GMA (Gesellschaft für Medizinische Ausbildung) [1] und die Basisstandards der WFME (World Federation for Medical Examination) [3] weisen zudem auf folgende Kriterien hin:

- Die Prüfungen müssen justiziabel sein.
- Das Prüfungsverfahren orientiert sich an Lernzielen und an der lernsteuernden Wirkung auf die Studierenden.
- Die verwendeten Prüfungsverfahren und die Grundsätze zum Bestehen der Prüfungen müssen bekannt gemacht werden.

Der Aufbau eines funktionierenden Evaluationssystems auf internationalem Niveau für Leistungskontrollen in den Universitäten wurde 2008 vom Wissenschaftsrat empfohlen. Die verwendeten Bewertungsinstrumente sollten die Lehrleistung verlässlich und transparent analysieren [http://www.wissenschaftsrat.de/download/archiv/8639-08.pdf, zuletzt abgerufen am 23.10.2015]. Dem steht gegenüber, dass die aktuell geltende Approbationsordnung für Zahnärzte aus dem Jahre 1955 keine Vorgaben zu den abzuhaltenden, studiumsbegleitenden Prüfungen beinhaltet [http://www.gesetze-im-internet.de/z_pro/BJNR000370955.html, zuletzt abgerufen am 23.10.2015].

Da im Zahnmedizinstudium verstärkt praktische Fertigkeiten vermittelt und somit auch geprüft werden, handelt es sich meistens um den Einsatz kompetenzorientierter Prüfungsformen, die auf der Miller-Pyramide mit „zeigt wie“ beziehungsweise „handelt“ charakterisiert werden können [4]. Aus diesem Kontext kommen vor allem die Prüfungsformen des OSCE (Objective Structured Clinical Examination) und OSPE (Objective Structured Practical Examination) in Frage [4].

Die Prüfungsform OSCE wurde im Jahr 1975 durch Harden eingeführt [5]. Zunächst für Prüfungen im Fach Medizin konzipiert, wird OSCE heute ebenfalls im Rahmen zahnmedizinischer Prüfungen angewandt. In einer Studie aus dem Jahr 1998 stellten Manogue und Brown [6] erstmals die Entwicklung und Ausführung von OSCE in der Zahnmedizin vor. Die Begriffe OSCE und OSPE werden in der Literatur meist äquivalent und somit nicht differenziert verwendet. Sowohl Natkin und Guild [7] als auch der AMEE (Association for Medical Education in Europe) Guide No. 81 Part I. [8] beschreiben OSPE - als eine Variation der OSCE - als Prüfungsmethode, um praktische Fertigkeiten und Wissen in einer nicht-klinischen Umgebung zu prüfen. Die Autoren Wani und Dalvi [9] stellten ergänzend fest, dass OSPE eine Prüfungsform sei, mit der sich die Stärken und Schwächen der studentischen, praktischen Fertigkeiten darstellen und überprüfen lassen. Sowohl Studierende als auch PrüferInnen bewerteten diese Prüfungsform als positiv und sinnvoll [10], [11], [1], [12], [13], [14]. In weiteren Studien, wie der Untersuchung von Smith et al. [15], Nayak et al. [16] und Abraham et al [12], bezeichneten die Studierenden sowohl OSCEs als auch OSPEs im Vergleich zu schriftlichen und mündlichen Prüfungen als gerechtere und weniger stressige Prüfungsformen und zogen die OSPE der „traditionellen“ Prüfungsform vor. Eine Untersuchung von Schoonheim-Klein et al. [17] konnte außerdem zeigen, dass speziell OSCEs im dentalen Kontext die Fähigkeiten im Bereich der klinischen Kompetenz, das Lernen selbst, sowie eine realistischere Selbsteinschätzung der Studierenden förderten. Zudem konnte die Studie von Nayak et al. [16] darstellen, dass durch OSPE neben den individuellen Kompetenzen eines jeden Studierenden, auch die praktische Demonstration von Fakten- und Handlungswissen, sowie das Lernverhalten positiv beeinflusst werden.

Für die OSCEs wurden Reliabilitätswerte zwischen 0.11 und 0.97 angegeben [18]. Die stark differierenden Ergebnisse erklären sich vor allem dadurch, dass die Parameter unter denen eine OSCE abgehalten wird (Stationsanzahl, PrüferInnenanzahl, Dauer der Prüfung, Art der Bewertungsmodi), starke Variationen aufweisen können.

Unabhängig von der Prüfungsart wird standardmäßig bei der Bewertung zwischen den Methoden der „glance and grade“ (= per Augenschein) und der Bewertung aufgrund definierter Kriterien unterschieden. Diese Methoden wurden auch im Kontext von zahnärztlichen Prüfungssettings evaluiert [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31]. Die meisten der oben genannten Studien konnten keine signifikanten Unterschiede zwischen der Augenschein- und der kriterienbasierten Methodik feststellen. Zudem fanden sie nicht in einer realen, sondern in einer artifiziellen Prüfungsumgebung statt.

Zu OSPE-Prüfungen, die wie bereits beschrieben im eigentlichen Sinne eine Variation der OSCE darstellen gibt es kaum Studien zur Einschätzung der weiter oben genannten Parameter. So ist es beispielsweise nicht erforscht, inwieweit die PrüferInnenanzahl und die Art der Bewertungsmethode das Ergebnis einer OSPE beeinflussen.

Vor diesem Hintergrund war es das Ziel der vorliegenden Studie, die Reliabilität einer realen OSPE-Semesterabschlussprüfung im Phantomkurs der Zahnerhaltungskunde in Frankfurt am Main unter Berücksichtigung unterschiedlicher Bewertungsmodi und PrüferInnenanzahl zu evaluieren.

Material und Methoden

Der Phantomkurs der Zahnerhaltungskunde lief jeweils über einen Zeitraum von einem Semester (16 Wochen). In dieser Zeit mussten die Studierenden praktische Arbeiten an verschiedenen Simulationsmodellen (an extrahierten humanen bzw. industriell hergestellten Kunststoffzähnen) absolvieren. Anhand von vorher definierten Behandlungsprotokollen wurden Schritt für Schritt verschiedene Therapiealternativen (beispielsweise Füllungen, Laborrestaurationen wie Inlays, endodontische Maßnahmen etc.) mit Unterstützung der Lehrenden eingeübt. Jeder Schritt wurde in einem sogenannten Testatheft von den betreuenden Lehrenden unterzeichnet, sobald die im Vorfeld definierten Kriterien erfüllt wurden. Der Lernprozess wurde mit formativem Feedback begleitet. Zum Abschluss des Kurses fand neben einer mündlichen Wissensüberprüfung auch eine summative OSPE statt. Letztgenannte wurde an der Simulationseinheit an sogenannten „Phantompatienten“ durchgeführt. Zwei Kunststoffmodelle (Ober- und Unterkiefer) wurden in einem „Phantomkopf“ bestehend aus jeweils 14 Kunststoffoberkiefer- und 14 Kunststoffunterkieferzähnen befestigt. Die OSPE bestand aus zwei Prüfungsteilen, der „Füllung“ (A) und dem „Inlay“ (B), durchgeführt an zwei verschiedenen Kunststoffzähnen der jeweiligen Modelle. Diese gliederten sich in insgesamt sechs „Untereinheiten“ (1. „Primärpräparation“; 2. „Unterfüllung und Sekundärpräparation“; 3. „Füllung“; 4. „Inlay“; 5. „Füllung gesamt“ und 6. „Gesamtnote“), die jeweils von den PrüferInnen benotet wurden (siehe Abbildung 1). Diese Untereinheiten entsprachen den Kriterien, auf deren Basis die Testate im Kursablauf von den Lehrenden erteilt wurden. Die Prüfer-Checkliste, die die oben genannte Aufzählung von Teilaspekten (Untereinheiten) beinhaltete, wurde im Vorfeld in vier aufeinanderfolgenden Semestern (SS 2008 bis WS 2009) im regulären Prüfungsszenario erprobt. Während der Erprobung erfolgte die Bewertung durch Inaugenscheinnahme der vorgegebenen Teilaspekte, alleine anhand von aus Sicht der PrüferInnen allgemeingültigen

Qualitätskriterien. Vergeben wurden Schulnoten von 1 bis 5 (1=sehr gut bis 5=mangelhaft).
Jeder Prüfer, jede Prüferin bewertete im realen Prüfungsszenario (Dauer: 3 h) jeden Studierenden. Das bedeutete, dass die PrüferInnen in einer festgelegten Reihenfolge die Arbeiten der Studierenden direkt am Arbeitsplatz (am Phantompatienten) während der laufenden Prüfung beurteilten. Die Studierenden meldeten den PrüferInnen durch Handzeichen, dass sie bereit waren, eine Untereinheit zur Bewertung vorzuzeigen. Die PrüferInnen tauschten während der laufenden OSPE untereinander keine Informationen über die jeweils vergebenen Noten aus. Nachdem die PrüferInnen unabhängig voneinander ihre jeweiligen Prüfer-Checklisten vervollständigt hatten, wurden in einer gemeinsamen Besprechungsrunde die Bewertungen diskutiert und festgelegt, welche Studierenden die Prüfung wiederholen sollten. Dies geschah nach dem Delphi-Prinzip [http://www.horx.com/zukunftsforschung/Docs/02-M-09-Delphi-Methode.pdf, zuletzt abgerufen am 23.10.2015].

Prüfungszenario der Studie

Die vorliegende Studie bezieht sich auf einen Zeitraum von zwei Semestern (SS 2010 = Gruppe I, SS 2012 = Gruppe II). Die Zusammensetzung der Studienpopulation ist in Tabelle 1 dargestellt. Die Einschlusskriterien lauteten:

- Studierende des 6. Semesters
- Teilnahme am Phantomkurs für Zahnerhaltungskunde
- Prüfungsfähigkeit vorhanden

Die Ausschlusskriterien waren wie folgt definiert:

- Studierende anderer Semester
- KursabbrecherInnen bzw. KurswiederholerInnen
- Prüfungsfähigkeit nicht gegeben

Der Unterschied in der jeweiligen Gruppengröße (I versus II) ergab sich aus der tatsächlichen Semestergröße, die großen Schwankungen unterlag und von den Ergebnissen des vorangestellten Physikums abhing. Eine zahlenmäßige Anpassung beider Gruppen war nicht durchführbar, da alle TeilnehmerInnen des Kurses laut Studienordnung an der Prüfung teilnehmen mussten. Die Festlegung der PrüferInnenanzahl erfolgte im Vorfeld dieser Studie beim Einreichen des Ethikantrages. Der Einsatz identischer PrüferInnen bei beiden Gruppen war aus Personalbesetzungsgründen in der Poliklinik nicht realisierbar.
In Gruppe I wurde ausschließlich eine Prüfer-Checkliste, wie in Abbildung 1 ersichtlich, angewendet. In Gruppe II verwendeten die PrüferInnen die identische Prüfer-Checkliste wie in Gruppe I, jedoch in Kombination mit einem detaillierten Dozentenmanual (siehe Abbildung 2). Dieser enthielt klar definierte Bewertungskriterien für die einzelnen Schulnoten.
Insgesamt nahmen fünf PrüferInnen (A-E), vier Frauen und ein Mann an der Studie teil. Die PrüferInnen waren ZahnärztInnen der Poliklinik für Zahnerhaltungskunde, hatten Erfahrung in der Lehre und der Bewertung von studentischen Arbeiten im Phantomkurs. Tabelle 2 zeigt deren Verteilung nach Anzahl und Geschlecht. PrüferIn A hatte im Jahr 1990, B 2007, C 2008, D 2010 und E 2011 das zahnärztliche Examen absolviert. Sie alle hatten Erfahrung in der Betreuung des Phantomkurses der Zahnerhaltungskunde. Lediglich A wies zusätzlich zu den anderen auch Erfahrung in der Betreuung von Patientenbehandlungskursen auf.
Die Prüfer-Checkliste entstand in Anlehnung an Themengebiete, die im laufenden Kurs und in den Lehrbüchern für Zahnerhaltungskunde standardmäßig inhaltlich abgebildet waren. Diese entsprachen zudem den in der Studie von Baumann [32] interdisziplinär zwischen vier Zentren (Universität Frankfurt, Freiburg, Leipzig und München) erhobenen Einheiten (Füllung, Inlay) und Untereinheiten, die im Fach Zahnerhaltungskunde als prüfungsrelevant definiert wurden. Dem für die Gruppe II beigefügten Manual konnten die Prüfer zusätzlich entnehmen, welche Bewertungskriterien erfüllt sein sollten, damit eine bestimmte Note vergeben werden konnte.

Train-the-Teacher

In jedem Semester fand eine 45-minütige „Train-the-Teacher-Veranstaltung“ statt. In diesem Seminar wurden die PrüferInnen durch praktische Übungen und theoretische Unterweisungen auf die Situationen in der OSPE und die Anwendung der Prüfer-Checkliste bzw. des Dozentenmanuals vorbereitet. So konnte im Vorfeld ein relativ hohes Maß an Standardisierung zwischen den PrüferInnen gewährleistet werden.

Statistik und Ethikantrag

Die Auswertung der Ergebnisse erfolgte nach der Generalisierbarkeitstheorie (G-Theorie) mit den Statistikprogrammen SAS 9.2 (SAS Institute Inc., Cary, USA, PROC MIXED) und R (Version 2.15, Package lme4). Die Varianz der erzielten Noten wird dabei auf die Einflussfaktoren (in der Terminologie der G-Theorie „Facetten“) „Studierender“ und „Untersucher“ sowie einer Messfehlerkomponente zurückgeführt (siehe Abbildung 3). Aus den Varianzanteilen der Facette Untersucher und der Fehlervarianz relativ zu dem der Facette „Studierender“ lässt sich die Messzuverlässigkeit der Bewertungen abschätzen. Der Generalisierbarkeitskoeffizient stellt dabei ein Analogon zur internen Konsistenz (Cronbachs alpha) dar. Im Unterschied zur üblichen Anwendung auf verschiedene Aufgaben wird er hier für verschiedene Prüfer verwendet. Die G-Theorie erlaubt eine Abschätzung der Messzuverlässigkeit bei Annahme einer anderen Zahl von Prüfern als in der tatsächlichen Untersuchung. Damit lassen sich die beiden Studien, bei denen eine unterschiedliche Zahl von PrüferInnen beteiligt waren, vergleichbar machen (analog zur Spearman-Brown-Formel, mit der eine Normierung der Reliabilität auf eine bestimmte Anzahl von Aufgaben möglich ist).
Analog hierzu wurden auch die einzelnen PrüferInnen (A-E) untereinander hinsichtlich des Parameters „Gesamt-

Tabelle 1: Zusammensetzung der Studienpopulation.

Gruppe	Anzahl Studierende	Frauen	Männer	Alter im Mittel (Jahre)
I	64	38	26	24,9
II	44	27	17	24,7

A.			**Füllung**
	I		Primärpräparation
		1.	Kontakt approximal
		2.	Boden plan und glatt
		3.	Winkel pulpaaxiale Wand
		4.	Breite und Tiefe
	II		Unterfüllung und Sekundärpräparation
		1.	UF glatt und Übergang gerundet
		2.	UF Lokalisation
		3.	UF Höhe
	III		Füllung
		1.	Randdichte
		2.	Kontaktpunkte
		3.	Okklusale Gestaltung
		4.	Glätte
			Füllung gesamt
B.			**Inlay**
	I.	1.	Kavitätenaußenränder
		2.	Kavitäteninnenwände
		3.	Breite und Tiefe
		4.	Glätte
		5.	Nachbarzahn
			Inlay gesamt
			GESAMTNOTE OSPE

Abbildung 1: Prüfer-Checkliste Gruppe I und Gruppe II mit den beiden Aufgabenstellungen A (Füllung) und B (Inlay). Die Abkürzung UF steht für Unterfüllung.

note OSPE“ evaluiert. Eine alle Parameter erfassende Subgruppenanalyse betreffend PrüferInnen A und B vervollständigte die statistische Analyse.

Ein Ethikantrag der monozentrischen Vergleichsstudie erhielt bei der Ethikkommission des Fachbereiches für Medizin der Goethe-Universität die Genehmigungsnummer 135/13.

Ergebnisse

Tabelle 3 zeigt die Ergebnisse der Reliabilitätsbestimmung aus Gruppe I bei Verwendung der Prüfer-Checkliste ohne Dozentenmanual. In dieser Gruppe wurden bei drei PrüferInnen nur für die zwei Kriterien „Kavitäteninnenwände“ und „Breite/Tiefe“ Cronbachs Alpha Werte unter 0,6 ermittelt.

Alle übrigen Untereinheiten konnten den für eine ausreichende Reliabilität geforderten Wert von 0,6 bzw. größer als 0,6 erreichen. Die Untereinheit „Nachbarzahn“ erzielte den Wert 1,0; was als idealer Reliabilitätswert anzusehen ist. Des Weiteren zeigt Tabelle 3 die Ergebnisse der Reliabilitätsbestimmung aus Gruppe II (Verwendung der Prüfer-Checkliste inklusive Dozentenmanual). Um eine Vergleichbarkeit der Generalisierbarkeitskoeffizienten in beiden Studien zu ermöglichen, wurden diese jeweils sowohl für eine Zahl von drei wie auch für vier PrüferInnen umgerechnet. So wurden für Studie I die Reliabilitätswerte

Primärpräparation

1. Kontakt approximal
1 alle 3 Flächen sondenspitzengroß aufgelöst
2 1 Fläche doppelte Sondenspitze aufgelöst
3 1 Fläche sondenspitzengroß und 2 Flächen doppelte Sondenspitze aufgelöst
4 alle 3 Flächen doppelte Sondenspitze aufgelöst
5 mind. 1 Fläche mehr als doppelte Sondenspitze bzw. mind. 1 Fläche nicht aufgelöst

2. Boden plan und glatt
1 optimal plan und glatt
2 +/- 5 Grad plan und glatt
3 plan oder glatt, +/- 5 Grad plan und rau (Glättung mit Arkansas wäre nötig)
4 +/- 10 Grad plan und glatt
5 > 10 Grad von plan

3. Winkel p.a. Wand
1 gebrochen und 90 Grad/leicht konisch
2 nicht gebrochen und 90 Grad/leicht konisch
3 nicht gebrochen und untersichgehend (5 Grad)
4 divergierend bis 5 Grad
5 untersichgehend/divergierend >5 Grad

4. Breite und Tiefe
1 optimal, 10er Birne versenkt, Breite 1/3 des bukkalen/oralen Höckerabstandes
2 Tiefe > als ¾ der Birne kann noch versenkt werden, Breite 1/3 des bukkalen/oralen Höckerabstandes
3 Breite und Tiefe weichen nicht mehr als 10% ab
4 Breite und Tiefe weichen nicht mehr als 15% ab
5 Füllung nicht möglich / P

Unterfüllung und Sekundärpräparation

1. UF glatt und Übergang gerundet
1 optimal plan und glatt, Übergang gebrochen
2 < 5 Grad Abweichung von plan, Kanten gerundet
3 maximal 5 Grad Abweichung von plan, Kanten gerundet
4 10 Grad Abweichung von glatt, Kanten gerundet
5 >10 Grad Abweichung von glatt, scharfe Kanten

2. UF Lokalisation
1 Schmelzränder frei, nur auf p.a. Wand und Kastenboden
2 Schmelzränder frei, 90% der p.a.Wand und Kastenboden von UF bedeckt
3 Schmelzränder frei, 80% der p.a.Wand und Kastenboden von UF bedeckt
4 Schmelzränder teilweise frei aber mehr als nur p.a. Wand und Kastenboden
5 p.a. Wand liegt >20% frei und/oder Phosphatzement auf Schmelzränder

3. UF Höhe
1 ¾ der Birne kann versenkt werden, optimal plan und glatt
2 ¾ der Birne kann versenkt werden, optimal plan, leichte Unebenheiten
3 ¾ der Birne kann versenkt werden, +/- 5 Grad plan und glatt
4 ¾ der Birne kann versenkt werden, +/- 10 Grad plan und glatt
5 < als ¾ der Birne können versenkt werden

Füllung

1. Randdichte
1 Füllung dicht, kein Spalt
2 Füllung annähernd dicht, Sonde bleibt minimal hängen
3 Füllung leicht über-/unterkonturiert
4 Spalt <0,5mm
5 Spalt deutlich tastbar

2. Kontaktpunkte
1 optimal (Matrizenband mit Abdruckspur)
2 minimal zu stark/zu schwach (Matrizenband ohne Abdruckspur)
3 Zahnseide hält
4 doppelte Zahnseide hält
5 Spalt sichtbar (doppelte Zahnseide hält nicht)/ Approximalkontakt verblockt

3. okklusale Gestaltung
1 anatomische Gestaltung
2 anatomisch ähnliche Gestaltung
3 Modellation min. 1 Hauptfissur und 2 Randleisten
4 Modellation min. 1 Hauptfissur
5 keine Konturierung erkennbar

4. Glätte
1 Politur erfolgt, Hochglanz erkennbar
2 Politur erfolgt, Glanz erkennbar
3 Politur erfolgt
4 Füllung nicht poliert
5 Poren, Einschlüsse, keine Konturierung mit rotierenden Instrumenten

Inlay

1. Kavitätenaußenränder
1 Dimension und approx. Auflösung optimal (doppelte Sondenspitze)
2 Dimension und approx. Auflösung annähernd optimal (sondenspitzengroß)
3 Präparation zu breit/zu schmal, approx. Auflösung min. sondenspitzengroß
4 Definierte Präparationsgrenze, approx. Auflösung min. sondenspitzengroß
5 Abformung und Inlay nicht möglich

2. Kavitäteninnenwände
1 konisch und abgerundet
2 Mängel in konisch bzw. abgerundet
3 Wände konisch aber nicht abgerundet
4 Kavität gerundet, Wände nur annähernd konisch (Inlay labortechnisch möglich)
5 Inlay nicht möglich

3. Breite und Tiefe
1 optimal, Tiefe min 1,5mm, Breite 1/3 des bukkalen/ oralen Höckerabstandes
2 annähernd optimal (Tiefe zwischen 1,5-2mm, Breite 1/3 des bukkalen/oralen Höckerabstandes)
3 Tiefe zwischen 1,5-2mm, Breite 10% Abweichung von 1/3 des bukkalen/oralen Höckerabstandes
4 Tiefe max. 2,5mm, Breite 10% Abweichung von 1/3 des bukkalen/oralen Höckerabstandes)
5 Tiefe weniger als 1,5mm, Isthmus zu eng/zu breit (>10% Abweichung von 1/3 des bukkalen/oralen Höckerabstandes)

5. Glätte
1 optimal plan und glatt
2 optimal plan und annähernd glatt
3 vereinzelt Unebenheiten
4 Riefen
5 keine Glättung erfolgt

5. Nachbarzahn
0 nicht anpräpariert
1 Note Abzug im Schmelz eine Seite anpräpariert
2 Noten Abzug im Schmelz beidseitig anpräpariert
durchgefallen einseitig im Dentin anpräpariert
durchgefallen beidseitig im Dentin anpräpariert

Gesamtnote

Abbildung 2: Dozentenmanual der Gruppe II mit Bewertungskriterien beider Aufgaben A und B. Die Abkürzung p.a. bedeutet Pulpaaxiale Wand, UF steht für Unterfüllung, min = mindestens, approx = approximal.

Tabelle 2: Daten der PrüferInnen A bis E, die die Gruppe 1 im SS 2010 (A, B, C) und die Gruppe 2 im SS 2012 (A, B, D, E) evaluiert haben (w = weiblich, m = männlich).

Gruppe	PrüferInnen	Anzahl der PrüferInnen	Geschlecht der PrüferInnen (w)	Geschlecht der PrüferInnen (m)
I	A, B, C	3	2	1
II	A, B, D, E	4	3	1

Note = Fähigkeit des Studierenden + Strenge des Untersuchers + Messfehler

Abbildung 3: Die Facetten der in der Studie durchgeführten Varianzanalyse.

Tabelle 3: Ergebnisse der Gruppe I und der Gruppe II. Es werden die korrespondierenden Reliabilitätswerte (= Cronbachs alpha) bei drei und vier PrüferInnen angegeben. In der Spalte „A vs B" werden die Ergebnisse der Subgruppenanalyse dargestellt. Die Kennzeichnung mit * bedeutet, dass abweichend vom Realprüfszenario eine Umrechnung auf eine andere PrüferInnenanzahl erfolgte (Abkürzung: UF = Unterfüllung, vs = versus).

				Gruppe I		Gruppe II	
				3	4*	3*	4
A.			**Füllung**				
	I		Primärpräparation	0,867	0,897	0,373	0,443
		1.	Kontakt approximal	0,642	0,705	0,311	0,376
		2.	Boden plan und glatt	0,722	0,776	0,504	0,575
		3.	Winkel pulpaaxiale Wand	0,696	0,753	0,513	0,584
		4.	Breite und Tiefe	0,84	0,875	0,382	0,451
	II		Unterfüllung und Sekundärpräparation	0,835	0,871	0,798	0,84
		1.	UF glatt und Übergang gerundet	0,725	0,779	0,69	0,748
		2.	UF Lokalisation	0,689	0,747	0,56	0,629
		3.	UF Höhe	0,691	0,749	0,753	0,803
	III		Füllung	0,796	0,839	0,716	0,77
		1.	Randdichte	0,663	0,724	0,666	0,726
		2.	Kontaktpunkte	0,655	0,717	0,347	0,415
		3.	Okklusale Gestaltung	0,718	0,772	0,371	0,44
		4.	Glätte	0,777	0,823	0,428	0,499
			Füllung gesamt	0,823	0,861	0,739	0,791
B.			**Inlay**	0,72	0,774	0,406	0,476
	I.	1.	Kavitätenaußenränder	0,678	0,738	0,102	0,132
		2.	Kavitäteninnenwände	0,548	0,617	0,311	0,376
		3.	Breite und Tiefe	0,302	0,365	0,339	0,406
		4.	Glätte	0,661	0,723	0,278	0,339
		5.	Nachbarzahn	1	1	0	0
			Inlay gesamt	0,72	0,774	0,406	0,476
			GESAMTNOTE OSPE	**0,839**	**0,874**	**0,633**	**0,697**

für vier PrüferInnen mit Hilfe der Spearman-Brown-Formel aus denen für drei PrüferInnen bestimmt bzw. für Studie II umgekehrt.

In Gruppe II zeigten die Ergebnisse für vier PrüferInnen hohe Varianzen in den ermittelten Cronbachs-Alpha-Werten. Für die 1. Untereinheit „Primärpräparation" und die dazugehörigen Kriterien („Kontaktpunkt approximal" bis „Breite/Tiefe") wurden Cronbachs-Alpha-Werte unter 0,6 ermittelt. Ebenso verhielt es sich für die Untereinheit „Füllung" und die dazugehörigen Kriterien „Kontaktpunkte", „okklusale Gestaltung" und „Glätte", für „Inlay gesamt" und die dazugehörigen Kriterien wie „Kavitätenaußenränder", „Kavitäteninnenwände", „Breite/Tiefe", „Glätte" und „Nachbarzahn". Die verbliebenen Untereinheiten und Kriterien konnten den für eine ausreichende Reliabilität geforderten Wert von 0,6 erreichen.

Beim Vergleich der einzelnen PrüferInnen untereinander hinsichtlich des Parameters „Gesamtnote OSPE" konnten im Sommersemester 2010 Korrelationskoeffizienten von 0,58 (A versus C), 0,64 (A versus B) und 0,68 (C versus B) ermittelt werden. Im Sommersemester 2012 fielen die korrespondierenden Werte niedriger aus (A versus B: 0,33; A versus E: 0,35; A versus D: 0,34; E versus D: 0,52; B versus D: 0,37 und E versus B: 0,35). Die Ergebnisse der Subgruppenanalyse (A versus B, die in beiden Studiengruppen eingesetzt wurden) sind Tabelle 3 zu entnehmen.

Diskussion

Limitationen

Eine Limitation der vorliegenden Studie liegt in der Art des gewählten Versuchsdesigns (historische Vergleichsgruppe), denn die Untersuchung wurde nicht innerhalb eines Semesters an einer Studienpopulation, sondern an zwei aufeinanderfolgenden Semestern an unterschiedlichen TeilnehmerInnen durchgeführt. Eine semesterinterne Teilung der summativen Prüfung aufgrund zweier verschiedener Bewertungsmodi wurde von der Ethikkom-

mission der Fakultät für unzulässig erklärt. Eine weitere Limitation sehen die Autoren darin, dass die PrüferInnen der beiden untersuchten Gruppen sowohl in der Anzahl als auch in der Team-Zusammensetzung ungleich waren. Lediglich zwei PrüferInnen (A und B) bewerteten vergleichend in beiden Studiengruppen. Zudem ist trotz der vor geschalteten Train-the-Teacher-Veranstaltungen von einem bestehenden Unterschied in der Lehrerfahrung auszugehen. Diese Variation ließ sich jedoch aus Personalgründen (Vertragsablaufszeiten) nicht homogenisieren. Die aufwendige statistische Analyse trägt dieser Limitation Rechnung und standardisiert die ungleiche Prüferzahl.

Bewertungsmodi

Über den Nutzen einer Prüfer-Checkliste in Bezug auf die Reliabilität einer Prüfung kann aus der derzeitigen wissenschaftlichen Datenlage kein eindeutiger Schluss gezogen werden. Nach aktuellem Forschungsstand gibt es nur wenige Studien, die sich mit verschiedenen Bewertungsmodi auseinander gesetzt haben [19], [20], [26], [28], [29], [33]. In der vorliegenden Studie konnten die besten Ergebnisse in Bezug auf eine hohe Reliabilität bei der Verwendung der Prüfer-Checkliste eruiert werden, bei der kein zusätzliches Dozentenmanual verwendet wurde. Zu einem vergleichbaren Ergebnis kam auch die Studie von Bazan und Seale [34], bei der eine ähnlich konzipierte Prüfer-Checkliste für eine Prüfungsbewertung zu einem vergleichbaren Reliabilitätswert für die Prüfung führte. Eine Erklärung hierfür könnte sein, dass der Differenzierungsgrad der Bewertungsvorgaben im Dozentenmanual möglicherweise zu detailliert war, um von den PrüferInnen während der praktischen Prüfung angewendet werden zu können und die Train-the-Teacher-Veranstaltung scheinbar nicht in der Lage war, einen vergleichbaren Bewertungsstandard bei den PrüferInnen zu setzen. Besonders deutlich wurde diese Problematik bei dem Teilschritt „Inlay: Nachbarzahn“, bei dem das sehr ausführliche Manual mit den definierten Unterpunkten zu einer massiven Verschlechterung der Cronbachs alpha-Werten führte. Dies steht auch im Einklang mit der Studie um die Autorengruppe Houpt und Kress [31], die ergab, dass, je enger der vorgegebene Bewertungsrahmen für ein Kriterium definiert war, umso eher Abweichungen in der Messgenauigkeit und Einschätzung der PrüferInnen auftraten. Beim direkten Vergleich der PrüferInnen A und B, die in beiden Semestern prüften, zeigte sich, dass die Verwendung des Manuals die im SS 2010 ermittelte mittlere Korrelation (0.68) auf einen Wert von 0.33 senkte. Trotzdem bleibt Klärungsbedarf, warum ausgerechnet dieser Teilschritt solch extreme Abweichungen bedingte. Möglicherweise bewirkte die Wortwahl der Zahnhartsubstanzdefinitionen (Schmelz und Dentin) eine Verwirrung seitens der PrüferInnen, denn die Prüfungsaufgabe wurde nicht an natürlichen Zähnen bestehend aus Schmelz und Dentin durchgeführt, sondern an Prüfungszähnen bestehend aus Kunststoff. Zukünftige Studien sollten die genaue Wortwahl der Manualparameter inhaltlich thematisieren.

Prüfungssetting

Im Unterschied zu bereits erwähnten Studien fand die Beurteilung durch die PrüferInnen in der vorliegenden Studie in einer realen Prüfungssituation statt. Als mögliche zukünftige Alternative bezüglich des Studiendesigns wäre hierfür denkbar, den PrüferInnen mehr Zeit für die Bewertung zu geben, was allerdings an der hier untersuchten Prüfung an der Universität Frankfurt am Main eine grundlegende Neukonzeption der Semesterabschlussprüfung erfordern würde. Bedenkt man, dass für die gesamte Prüfung drei Stunden angesetzt wurden, und dass die einzelnen Schritte gleichzeitig bei durchschnittlich n=22 Studierenden adhoc durch die PrüferIn beurteilt wurden, so wäre ein längeres Verweilen bei der Beurteilung nur schwierig zu realisieren. Es stellt sich die Frage, warum während des realen OSPE-Prüfungszenarios ein solcher Aufwand betrieben wird und warum die einzelnen Schritte nicht nach der Prüfung gemeinsam mit allen PrüferInnen beurteilt werden können. Dies liegt daran, dass viele Einzelschritte während der Prüfung durch den darauffolgenden Schritt nicht mehr beurteilbar, da nicht mehr sichtbar sind. Beispielsweise ist der Schritt der „Primärpräparation“ nach der „Unterfüllung/ Sekundärpräparation“ nicht mehr beurteilbar, weil Ersterer nach dem Legen einer Unterfüllung teilweise verdeckt ist. So verhält es sich mit allen Teilschritten, so dass am Ende des Prüfungsabschnittes „Füllung“ nur noch der endgültig resultierende Schritt beurteilbar bliebe.
Dieses Vorgehen steht im großen Gegensatz zu allen bisher publizierten OSPE-Prüfungen, bei denen in der Regel die Einzelschritte auch nach der Prüfung noch sichtbar und beurteilbar waren. Verglichen mit den Studien von Goepferd und Kerber [26], Vann et al. [28] und Scheutzel [33] ergibt sich ein deutlicher Unterschied, da für die dort untersuchten Prüfungen der ähnlich komplexe Bewertungsbogen unter günstigeren Zeitvoraussetzungen angewendet werden konnte. Dies könnte die unterschiedlichen Ergebnisse zwischen der hier durchgeführten Untersuchung und den zuvor erwähnten Studien erklären.

Train-the-Teacher

OSCE-basierte Prüfungen weisen in Analogie zu den bereits weiter oben erwähnten Vorteilen auch einige Nachteile auf. Nach Miller [4], [35] haben Erfahrungen gezeigt, dass OSCE besonders trainings- und zeitaufwendig ist und nach Nayak et al. [16] einer intensiven Planung und Teamarbeit bedarf. In der Regel benötigen die eingesetzten PrüferInnen ein intensives, systematisches Training, um die Anforderungen an Reliabilität und Validität einer OSCE-Prüfung zu erfüllen [35]. OSCE ist folglich, im Vergleich zu anderen Prüfungsarten wie Multiple-Choice-Fragen oder mündliche Prüfungen, zeit- und vor allem kostenintensiv [8], [35], [36]. Auch im Rahmen der hier vorliegenden Studie wurde eine zeitintensive Vorbereitung der PrüferInnen in einer Train-the-Teacher-Veranstaltung durchgeführt. Dadurch mussten im klinischen und organisatorischen Arbeitsablauf in der Abteilung für Zahner-

haltungskunde personelle und räumliche Ressourcen und damit auch finanzielle Mittel gebunden werden. Die Dauer einer Vorlesungseinheit (45 min.) war hierfür realistisch gewählt und konnte von allen PrüferInnen wahrgenommen werden. Es stellt sich jedoch die Frage, wie lang eine Vorbereitung effektiv ausfallen muss um Erfahrungsunterschiede bei gemischten Teams im Vorfeld homogenisieren zu können. Im SS 2010 zeigten die drei PrüferInnen untereinander eine mittlere Korrelation zwischen 0.58 und 0.68. Im SS 2012 führte die identisch lang durchgeführte Train-the-Teacher-Veranstaltung bei den vier PrüferInnen zu Korrelationswerten zwischen 0.33 und 0.52. Hier kann vermutet werden, dass im Falle des angewendeten Manuals die Train-the-Teacher-Veranstaltung nicht effektiv eingesetzt wurde.

PrüferInnen

Bei der Reliabilitätswertung spielen nach der heutigen Datenlage die PrüferInnen eine wichtige Rolle. Bisher gibt es allerdings keine uns bekannten wissenschaftlichen Untersuchungen, die eine Aussage treffen, wie hoch die Mindestanzahl an PrüferInnenn für eine OSPE sein sollte. In der hier vorliegenden Studie konnte mit drei PrüferInnen eine ausreichend hohe Reliabilität in Kombination mit Check-Listen erzielt werden. Nach Ergebnissen dieser Untersuchung kann der Reliabilitätswert allerdings durch eine höhere Prüferzahl weiter gesteigert werden. Diese Steigerung der Reliabilitätswerte fällt im Verhältnis zu der PrüferInnenanzahl jedoch gering aus. Darüber hinaus würde eine weitere Erhöhung der PrüferInnenanzahl zu einem gesteigerten Aufwand hinsichtlich Organisation und finanziellen Kosten führen.
In diesem Zusammenhang muss kritisch erwähnt werden, dass aus den vorliegenden Daten keine generelle Empfehlung für andere Standorte bezüglich der PrüferInnenanzahl abgegeben werden kann, da die Möglichkeit, drei bis vier lang erfahrene PrüferInnen für eine OSPE-Prüfung zur Verfügung zu haben, für viele Standorte durchaus nicht die Regelsituation darstellt bzw. nicht realisierbar ist. Zu einem ähnlichen Ergebnis in Bezug auf den gesteigerten Aufwand hinsichtlich Organisation bei OSCE-Prüfungen kamen auch die Autorengruppen um Nikendei und Jünger [37] bzw. Norcini et al. [38]. Natkin und Guild [39] konnten in ihrer Arbeit durch eine systematische Vorbereitung der PrüferInnen eine deutliche Reliabilitätssteigerung nachweisen. Ähnliche Ergebnisse stellte auch Dhuru [25] vor, in dessen Arbeit BewerterInnen mit langjähriger Berufserfahrung und bei Verwendung eines Bewertungsbogens die reliabelsten Prüfungsergebnisse erzielten. Dies kann in der vorliegenden Studie lediglich bei der Verwendung der Checkliste bestätigt werden, denn die zwei Prüferinnen mit der längsten Erfahrung wiesen im Falle des verwendeten Manuals lediglich schwache Korrelationen auf. Die Checkliste scheint, wie in dieser Untersuchung deutlich wird, in der Lage zu sein, die Reliabilität weiter zu erhöhen beziehungsweise mangelnde Prüfungserfahrung aufseiten der Bewertenden zu kompensieren. Dagegen konnte in der Untersuchung von Houpt und Kress [31] die Reliabilität nicht bei allen Bewertungskriterien gesteigert werden. Somit scheint es nach Meinung der Autoren, dass Train-the-Teacher-Veranstaltungen alleine nicht in der Lage sind, die Interrater-Reliabilität signifikant zu erhöhen. Derartige Trainingsveranstaltungen hatten den größten Effekt bei „Non-Expert"-Prüfern, dagegen relativ geringen Einfluss bei erfahrenen BewerterInnen [31]. Dies kann auch von unserer Untersuchung bestätigt werden.

Prüfungsaufgaben

Die Anzahl der in dieser Studie definierten Prüfungsaufgaben, die man häufig in der Literatur mit dem Begriff der „Stationen" gleichsetzt, sollte kritisch hinterfragt werden. Im vorliegenden Fall waren es zwar nur zwei getrennte Aufgaben (A. Füllung und B. Inlay), jedoch insgesamt 22 Bewertungen, die man als BewerterIn pro Studierenden in und während der Prüfung abgab. Es geht letztlich um die Definition des Begriffes „Station" in Zusammenhang mit einer OSPE, was evidenzbasiert aus der Literatur nicht abzuleiten ist. Es bleibt zudem kritisch anzumerken, dass ein Wert von 0,6 für Cronbachs alpha lediglich einen „ausreichenden" Charakter besitzt. Es ist ebenfalls zu hinterfragen, wie valide eine Prüfung dann überhaupt ist und ob sie sich für eine summative Prüfung eignet. Vor diesem Hintergrund lässt sich die Variante II aus unserer Sicht für „high stakes" Examina nach der vorliegenden Datenlage nicht empfehlen.

Schlussfolgerung

Aus der vorliegenden Studie ergeben sich folgende Schlussfolgerungen hinsichtlich der Frage, wie eine OSPE in der zahnmedizinischen Lehre im Phantomkurs der Zahnerhaltungskunde möglichst reliabel gestaltet werden kann:

- Eine Prüfer-Checkliste ohne Dozentenmanual ergab eine höhere Interrater-Reliabilität im Rahmen der durchgeführten OSPE.
- Die Bewertung der studentischen Prüfungsleistungen im Rahmen der OSPE sollte nach Möglichkeit durch mindestens drei PrüferInnen vorgenommen werden.

Danksagung

Die Autoren bedanken sich bei den Studierenden des 6. Semesters im Fach Zahnerhaltungskunde und bei den zahnärztlichen KursassistentInnen, die bei der Bewertung der OSPE ihren Beitrag geleistet haben.

Interessenkonflikt

Die Autoren erklären, dass sie keinen Interessenkonflikt im Zusammenhang mit diesem Artikel haben.

Literatur

1. Gesellschaft für Medizinische Ausbildung, Kompetenzzentrum Prüfungen Baden-Württemberg, Fischer MR. Leitlinie für Fakultätsinterne Leistungsnachweise während des Medizinstudiums: Ein Positionspapier des GMA-Ausschusses Prüfungen und des Kompetenzzentrums Prüfungen Baden-Württemberg. GMS Z Med Ausbild. 2008;25(1):Doc74. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2008-25/zma000558.shtml
2. Taylor CL, Grey NJ, Satterthwaite JD. A comparison of grades awarded by peer assessment, faculty and a digital scanning device in a pre-clinical operative skills course. Eur J Dent Educ. 2013;17(1):16-21. DOI: 10.1111/j.1600-0579.2012.00752.x
3. World Federation for Medical Education. Basic Medical Education The 2012 Report. Copenhagen: WFME Office; 2012.
4. Miller GE. The assessment of clinical skills/competence/performance. Acad Med.1990;65:S63-67. DOI: 10.1097/00001888-199009000-00045
5. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J. 1975;1:447-451. DOI: 10.1136/bmj.1.5955.447
6. Manogue M, Brown G. Developing and implementing an OSCE in dentistry. Eur J Dent Educ.1998;2(2):51-57. DOI: 10.1111/j.1600-0579.1998.tb00039.x
7. Natkin E, Guild RE. Evaluation of preclinical laboratory performance: a systematic study. J Dent Educ.1967;31(2):152-161.
8. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. Med Teach. 2013;35(9):e1437-1446. DOI: 10.3109/0142159X.2013.818634
9. Wani P, Dalvi V. Objective Structured Practical Examination vs Traditional Clinical Examination in Human Physiology: Students perception. Int J Med Sci Public Health. 2013;2(3):522–547. DOI: 10.5455/ijmsph.2013.080320133
10. Schoonheim-Klein M, Muijtjens A, Muijtens A, Habets L, Manogue M, van der Vleuten C, Hoogstraten J, Van der Velden U. On the reliability of a dental OSCE, using SEM: effect of different days. Eur J Dent Educ. 2008;12(3):131–137. DOI: 10.1111/j.1600-0579.2008.00507.x
11. Hofer M, Jansen M, Soboll S. Potential improvements in medical education as retrospectively evaluated by candidates for specialist examinations. Dtsch Med Wochenschr. 2006;131(8):373–378. DOI: 10.1055/s-2006-932527
12. Abraham RR, Raghavendra R, Surekha K, Asha K. A trial of the objective structured practical examination in physiology at Melaka Manipal Medical College. India. Adv Physiol Educ. 2009;33(1):21–23. DOI: 10.1152/advan.90108.2008
13. Adome RO, Kitutu F. Creating an OSCE/OSPE in a resource-limited setting. Med Educ. 2008;42(5):525–526. DOI: 10.1111/j.1365-2923.2008.03045.x
14. Davenport ES, Davis JE, Cushing AM, Holsgrove GJ. An innovation in the assessment of future dentists. Br Dent J. 1998;184(4):192–195.
15. Smith LJ, Price DA, Houston IB. Objective structured clinical examination compared with other forms of student assessment. Arch Dis Child. 1984;59:1173-1176. DOI: 10.1136/adc.59.12.1173
16. Nayak V, Bairy KL, Adiga S, Shenoy S, Magazine BC, Amberkar M, Kumari M. OSPE in Pharmacology: Comparison with the conventional Method and Students' Perspective Towards. Br Biomed Bull. 2014;2(1):218-222.
17. Schoonheim-Klein ME, Habets LL, Aartman IH, van der Vleuten CP, Hoogstraten J, van der Velden U. Implementing an Objective Structured Clinical Examination (OSCE) in dental education: effects on students' learning strategies. Eur J Dent Educ. 2006;10(4):226-235. DOI: 10.1111/j.1600-0579.2006.00421.x
18. Chenot JF, Ehrhardt M. Objective structured clinical examination (OSCE) in der medizinischen Ausbildung: Eine Alternative zur Klausur. Z Allg Med. 2003;79(2):437-442.
19. Sharaf AA, AbdelAziz AM, El Meligy OA. Intra- and inter-examiner variability in evaluating preclinical pediatric dentistry operative procedures. J Dent Educ. 2007;71(4):540-544.
20. Kellersmann CT. Zur Reliabilität der Beurteilung vorklinischer Phantomarbeiten bei Einsatz eines strukturierten Bewertungsbogens. Inaugural-Dissertation. Münster: Westfälischer Wilhelms-Universität Münster; 2007.
21. Lilley JD, ten Bruggen Cate HJ, Holloway PJ, Holt JK, Start KB. Reliability of practical tests in operative dentistry. Br Dent J. 1968;125(5):194-197.
22. Fuller JL. The effects of training and criterion models on interjudge reliability. J Dent Educ. 1972;36(4):19-22.
23. Hinkelman KW, Long NK. Method for decreasing subjective evaluation in preclinical restorative dentistry. J Dent Educ. 1973;37(9):13-18.
24. Gaines WG, Bruggers H, Rasmussen RH. Reliability of ratings in preclinical fixed prosthodontics: effect of objective scaling. J Dent Educ. 1974;38(12):672-675.
25. Dhuru VB, Rypel TS, Johnston WM. Criterion-oriented grading system for preclinical operative dentistry laboratory course. J Dent Educ.1978;42(9):528-531.
26. Goepferd SJ, Kerber PE. A comparison of two methods for evaluating primary class II cavity preparations. J Dent Educ. 1980;44(9):537-542.
27. Feil PH. An analysis of the reliability of a laboratory evaluation system. J Dent Educ. 1982;46(8):489-494.
28. Vann WF, Machen JB, Hounshell PB. Effects of criteria and checklists on reliability in preclinical evaluation. J Dent Educ. 1983;47(10):671-675.
29. Bedi R, Lo E, King NM, Chan T. The effect of pictorial criteria upon the reliability of assessments of cavity preparations. J Dent. 1987;15(5):222-224. DOI: 10.1016/0300-5712(87)90116-3
30. Jenkins SM, Dummer PM, Gilmour AS, Edmunds DH, Hicks R, Ash P. Evaluating undergraduate preclinical operative skill; use of a glance and grade marking system. J Dent. 1998;26(6):679-684. DOI: 10.1016/S0300-5712(97)00033-X
31. Houpt MI, Kress G. Accuracy of measurement of clinical performance in dentistry. J Dent Educ. 1973;37(7):34-46.
32. Baumann MP. Evaluation von Bewertungskriterien für praktische Studentenarbeiten im Vergleich zur Bewertung per Augenschein. Inaugural-Dissertation. München: Medizinischen Fakultät der Ludwig-Maximilians-Universität München; 2015.
33. Scheutzel P. Einfluss des Bewertungssystems auf Objektivität und Reliabilität der Benotung zahnmedizinischer Studentenarbeiten am Phantompatienten. GMS Z Med Ausbild. 2007;24(1):Doc67. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2007-24/zma000361.shtml
34. Bazan MT, Seale NS. A technique for immediate evaluation of preclinical exercises. J Dent Educ. 1982;46(12):726-728.
35. Barman A. Critiques on the Objective Structured Clinical Examination. Ann Acad Med Singapore. 2005;34(8):478-482.
36. Boursicot K, Ware J, Hazlett C. Objective Structured Clinical Examination Objective Structured Practical Examination. Med Educ. 1979;31:41-54.

37. Nikendei C, Jünger J. OSCE-praktische Tipps zur Implementierung einer klinisch-praktischen Prüfung. GMS Z Med Ausbild. 2006;23(3):Doc47. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2006-23/zma000266.shtml

38. Norcini JJ, Maihoff NA, Day SC, Benson JA. Trends in medical knowledge as assessed by the certifying examination in internal medicine. JAMA. 1989;262(17):2402–2404. DOI: 10.1001/jama.1989.03430170064029

39. Natkin E, Guild RE. Evaluation of preclinical laboratory performance: a systematic study. J Dent Educ. 1967;31(2):152-161.

Korrespondenzadresse:
PD Dr. med. dent. Susanne Gerhardt-Szép, MME
Goethe-Universität Frankfurt am Main, Carolinum Zahnärztliches Universitäts-Institut gGmbH, Poliklinik für Zahnerhaltungskunde, 60596 Frankfurt/Main, Deutschland, Tel.: +49 (0)69/6301-7505, Fax: +49 (0)69/6301-3841
s.szep@em.uni-freiburg.de

Bitte zitieren als
Schmitt L, Möltner A, Rüttermann S, Gerhardt-Szép S. Study on the Interrater Reliability of an OSPE (Objective Structured Practical Examination) – Subject to the Evaluation Mode in the Phantom Course of Operative Dentistry. GMS J Med Educ. 2016;33(4):Doc61. DOI: 10.3205/zma001060, URN: urn:nbn:de:0183-zma0010608

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/zma/2016-33/zma001060.shtml

Eingereicht: *23.10.2015*
Überarbeitet: *01.04.2016*
Angenommen: *03.06.2016*
Veröffentlicht: *15.08.2016*