Standardized examinees: development of a new tool to evaluate factors influencing OSCE scores and to train examiners

Petra Zimmermann[1]
Martina Kadmon[2]

1 Ludwig-Maximilians-Universität München, Klinikum der Universität, Klinik für Allgemein-, Viszeral- und Transplantationschirurgie, München, Germany

2 Universität Augsburg, Medizinische Fakultät, Gründungsdekanat, Augsburg, Germany

Abstract

Introduction: The Objective Structured Clinical Examination (OSCE) is an established format for practical clinical assessments at most medical schools and discussion is underway in Germany to make it part of future state medical exams. Examiner behavior that influences assessment results is described. Erroneous assessments of student performance can result, for instance, from systematic leniency, inconsistent grading, halo effects, and even a lack of differentiation between the tasks to be performed over the entire grading scale. The aim of this study was to develop a quality assurance tool that can monitor factors influencing grading in a real OSCE and enable targeted training of examiners.

Material, Methods and Students: Twelve students at the Medical Faculty of the University of Heidelberg were each trained to perform a defined task for a particular surgical OSCE station. Definitions were set and operationalized for an excellent and a borderline performance. In a simulated OSCE during the first part of the study, the standardized student performances were assessed and graded by different examiners three times in succession; video recordings were made. Quantitative and qualitative analysis of the videos was also undertaken by the study coordinator.

In the second part of the study, the videos were used to investigate the examiners' acceptance of standardized examinees and to analyze potential influences on scoring that stemmed from the examiners' experience.

Results: In the first part of the study, the OSCE scores and subsequent video analysis showed that standardization for defined performance levels at different OSCE stations is generally possible. Individual deviations from the prescribed examinee responses were observed and occurred primarily with increased complexity of OSCE station content.

In the second part of the study, inexperienced examiners assessed a borderline performance significantly lower than their experienced colleagues (13.50 vs. 15.15, $p=0.035$). No difference was seen in the evaluation of the excellent examinees. Both groups of examiners graded the item "ocial competence" – despite identical standardization – significantly lower for examinees with borderline performances than for excellent examinees (4.13 vs. 4.80, $p<0.001$).

Conclusion: Standardization of examinees for previously defined performance levels is possible, making a new tool available in future not only for OSCE quality assurance, but also for training examiners. Detailed preparation of the OSCE checklists and intensive training of the examinees are essential.

This new tool takes on a special importance if standardized OSCEs are integrated into state medical exams and, as such, become high-stakes assessments.

Keywords: OSCE, OSPE, examiner training, quality assurance, standardized examinees

Introduction

The Objective Structured Clinical Examination (OSCE) is an established assessment format at most medical schools and is especially suited for evaluating practical clinical skills [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15]. An AMEE guideline defines binding standards and metrics for ensuring the quality of OSCEs [9]. The creation of blueprints for both the exam content and the exam format is recommended for all required assessments. A blueprint mapping out exam content and the corresponding stations for the respective subject areas should also form the basis of an OSCE. Based on the blueprint, checklists are created and critically reviewed, and standards are set for performance expectations. A good reliability and inter-rater reliability can be achieved through a sufficient number of OSCE stations, regular standard setting, adaption of the checklists, and regular examiner training. Test statistical analysis of the results should be used to detect problems with the checklists or examiners and to minimize problems by regularly repeating the process described above [9], [15], [16], [17], [18].

Many studies analyze potential factors that influence OSCE scores. These factors take on particular importance when the assessment format is used for a high-stakes exam, as is currently being discussed in Germany in regard to the state medical examinations [19]. Harasym et al. were able to show that stringency or leniency on the part of the examiners can lead to scores that are systematically too high or too low [13]. The student's performance level also appears to influence the reliability of the scores given by examiners. Byrne et al. describe that a good student performance was evaluated with more precision than a borderline performance [4]. Yeates et al. determined in several studies that a good performance was graded higher if the performance immediately prior to it was a poor one [7], [20]. At the same time, a borderline performance was assessed lower if the examiner had observed a good performance immediately before. In addition, the effects on grading as a result of halo effects and a lack of differentiation on the entire grading scale have also been described [21]. Schleicher et al. were able to show in a study encompassing multiple medical schools that student performances were assessed differently by local and central examiners. Simultaneously, a trend was seen toward different grading behavior depending on the genders of the examiners and examinees [22]. All previous studies on potential influencing factors and on quality assurance of the test format are based on analyses of results from live observations or videos of OSCEs. Although these analyses are based on OSCEs that, in general, were preceded by a standardized briefing of the examiners, they were, however, subject to potential influences stemming from the examinees and were not standardized, so that, ultimately, examiner characteristics could not be fully isolated for analysis.

A suitable tool does not yet exist to simulate potential influences stemming from the examinee for direct analysis of such influences on examiner behavior and exam results. At the same time, no suitable tool has been available to train examiners in a targeted manner regarding the potential limitations concerning the reliability of grading OSCE performance.

Simulated patients are now an integral part of medical education and medical assessments. They offer an opportunity to practice physician-patient interactions in a safe environment and these patients can play an assigned role in a standardized manner. At the same time, it is possible to vary the individual parameters, e.g. the simulated patient's reaction or the extent of the disease, to simulate different situations for students [23], [24], [25]. Based on the concept of simulated patients, it was our aim to transfer this concept of standardization to student performance on an OSCE. In the first part of this study, we investigated the possibility of training students to reproduce a defined performance on an OSCE. In the second part, we used the video recordings from the first part to analyze the influence of examiner experience on the grades they assigned for the performances and to evaluate the basic acceptance of standardized examinees by examiners.

As a result, there is a new tool for OSCE quality assurance that also enables the identification of individual factors influencing assessment and the targeted training of examiners in the future.

Material, methods and students

Twelve students were each trained to perform in a standardized manner at three different stations of the OSCE on surgery at the Medical Faculty of the University of Heidelberg. Per station, two students were taught to give a standardized excellent performance and two students to give a standardized borderline performance; there was one female and one male student for each performance level. A student who had been prepared to give an excellent performance at the OSCE abdominal examination station was unable to participate on short notice for health reasons.

The score for an excellent performance was defined as the maximum number of possible points on the checklist, minus no more than two points; a borderline performance was the required minimum number of points to pass, plus or minus one point (minimal competency).

The lowest passing score for the entire OSCE is the sum total of all minimum competencies on Heidelberg's surgical OSCE.

Figure 1 illustrates the study design; figure 1A describes the first part of the study and figure 1B the second.

OSCE checklists

Three checklists were selected whose use was already well established in the surgical OSCE and which had undergone repeated internal review. These checklists were for the following OSCE stations:

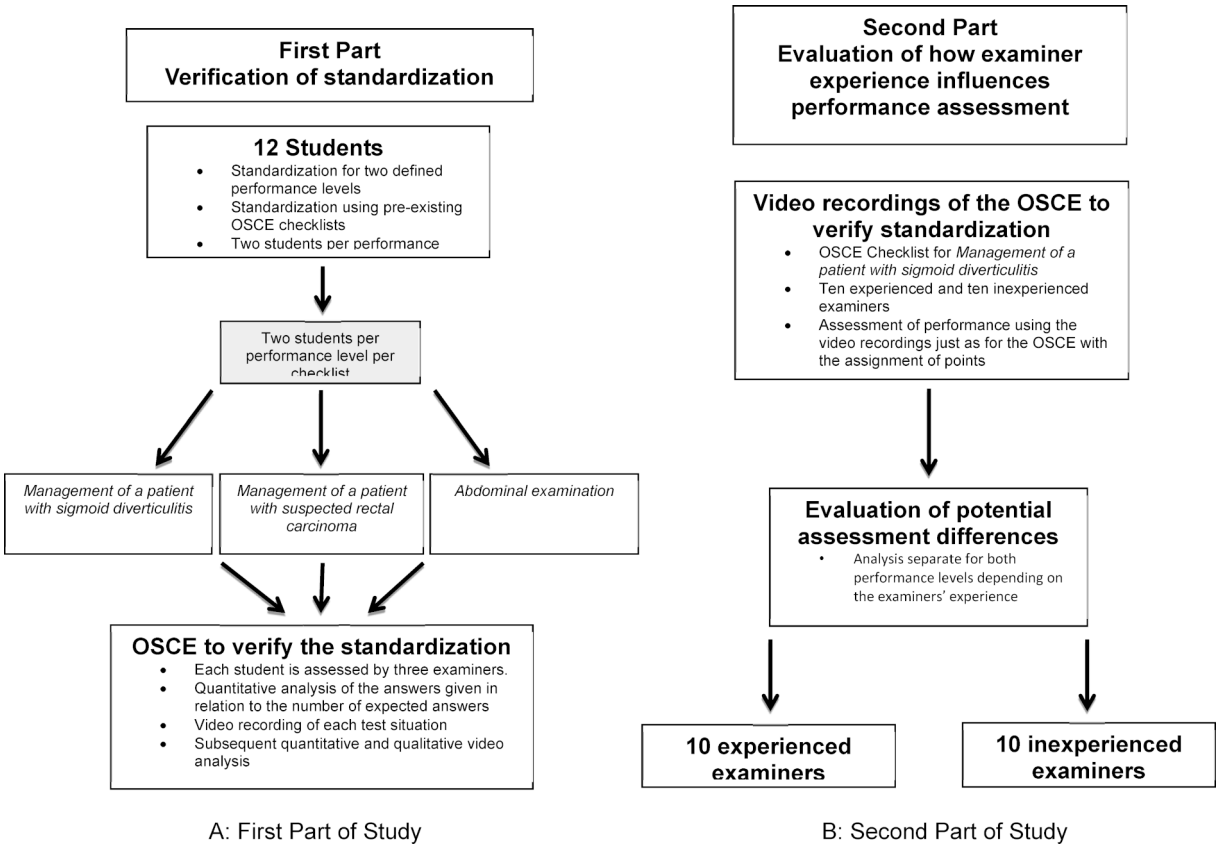

Figure 1: Schematic illustration of the study design with the first and second parts of the study

- *Management of a patient with sigmoid diverticulitis;*
- *Management of a patient with suspected rectal carcinoma;*
- *Abdominal examination.*

All of the checklists had a minimum of 0 and a maximum of 25 points. Each checklist consisted of five items, for which a maximum of five points each could be given. Each item covered a different number of required answers.

Minimal competency was defined as the number of points on a checklist necessary to pass. This also defines the minimum expectancy for each station based on the checklists and is routinely reviewed and defined by way of internal standard setting. The minimal competency for the checklists used was 17 points.

The maximum length of time for the exam was nine minutes per checklist; one minute was given to move between stations. The checklists also listed the grading subcategories (e.g. anamnesis, clinical exam, etc.) and the relevant individual items for assigning points:

- 5 points: all items completed without assistance;
- 3 points: all items completed in full with assistance from the examiner;
- 1 point: items were not fully completed despite assistance from the examiner.

It is clear for each graded category whether points should be given globally for overall impression or on the basis of answers to individual items.

Each checklist contains a brief case vignette, a task for each individual item, and the expected answers. Possible questions asked along the way by the examiner are not predefined.

The station checklists for *sigmoid diverticulitis and rectal carcinoma* cover the taking of the standardized patient's history (item 1), the determination of differential diagnoses based on the case history details (item 2), the decision which suitable diagnostics should be done in the actual situation (item 3), and for the sigmoid diverticulitis station, the description of a CT image from the patient case. Item 4 on both checklists covers the interaction with the standard patient regarding further diagnostic/therapeutic measures. Item 5 evaluates social competence. This also includes the extent to which the students adequately introduce themselves to the patients, how they behave toward the patients, for instance, if they are able to keep eye contact.

The checklist for the abdominal examination station covers the sequential steps to examine a patient with lower abdominal pain on the right side (item 1), checking for signs of peritonitis (item 2), explaining the performance of a digital rectal exam and the findings (item 3), examining the liver (item 4), and examining the spleen (item 5).

Modification of the OSCE checklists

To standardize the performance of the standardized examinees and to verify that this performance can be repro-

duced repeatedly, two new versions of the existing checklists used for the surgical OSCE were generated.

Checklists to standardize the examinees

To standardize the examinees, all of the checklists were operationalized in detail. For the two defined levels of performance, it was determined for each possible answer to a checklist item, whether the examinees should respond with a certain answer or not. In another field it was noted how the examinees should conduct themselves when asked a particular question, e.g. to answer hesitantly or only when prompted (see figure 2).

Checklists for evaluating performance

For the examiner to grade the performance, the evaluation part of the OSCE checklists was modified so that the examiner could note for each possible answer to each task whether or not that answer had been given (see figure 3). To eliminate potential systematic differences in assessment by the examiners, we did not carry out the standardization of the evaluation at the performance level using a global point value for each item, as is done in a real OSCE. A section was added at the end of the checklist in which the examiner was meant to evaluate the performance level using a global grading scale (poor, mediocre, very good) and the authenticity. Concerning the latter, the examiners were asked to evaluate the extent to which they doubted having a real examinee in front of them.
The examiners received the standardized assessment instructions for the surgical OSCE. However, they were instructed not to give any points for the individual items, but rather to tick each possible answer and indicate whether or not it had been given. The examiners were informed only after the OSCE that a standardization of student performance had been undertaken.

Standardized students

All 12 students had already completed the Surgery Block and taken the OSCE on surgery. The Surgical Block lasts for one semester and covers the subjects of visceral, vascular, thoracic and heart surgery, urology, orthopedics and trauma surgery, hand and plastic surgery, along with anesthesiology and emergency medicine. Lectures and seminars on pathology and radiology are integrated into the individual subject disciplines.
The students were given the checklists for training. The roles and the expected answers on the modified checklists were discussed in detail with each student. After two weeks to learn the checklists and roles, the test situation was simulated between the students and the study coordinator and corrections were made. As this was done, general challenges were discussed at first and then simulated in real-time as a test situation. Feedback was then given on the necessary changes.

First part of the study

Process of standardization

In the first part of this study (see figure 1, left A), standardization was carried out in a simulated OSCE that was held under real test conditions (time, time to change stations, etc.). The standardized examinees played their roles three times for three different examiners (one male examiner and two female examiners) and were recorded on video. In an additional second step, the videos were analyzed quantitatively and quantitatively by the study coordinator using the modified checklists so that there were six evaluations for each student.
When carrying out the quantitative analysis, the deviations were counted based on the prescribed answers that were supposed to have been given. The instances in which too many or too few answers were given were counted in relation to the correct number of expected answers. The mean percentages of the deviations were calculated for all OSCE run-throughs (3 test situations) and for the quantitative analysis from the subsequent video analysis.
When carrying out the qualitative analysis, the overall impression was evaluated first: The examinee appeared to be authentic (yes/no) and stayed in the standardized role. The following aspects were also evaluated:

- Conduct of the examinee when giving answers (appears confident, unconfident, tends to recite lists);
- Reaction of the examinee to the examiner's behavior/questions (stays in the role, deviates from the prescribed answers, lets him or herself be forced to give answers);
- Reaction of the examinee to the standard patient's behavior/questions (stays in the role, deviates from the prescribed answers, lets him or herself be forced to give answers);
- Conduct of the examiners;
- Conduct of the standard patients.

The study coordinator shared responsibility for the organization of the Surgery Block and had acted as an examiner more than 20 times in surgical OSCEs. In addition, she was experienced in the writing of OSCE checklists and exam questions. This study was carried out within the scope of her master's thesis to attain a Master of Medical Education in Germany (MME-D).

Second part of the study

Analysis of the influence of examiner experience on performance assessment

In the second part of the study (see figure 1, right B), the videos were used to investigate the influence of examiner experience on performance assessment and their acceptance of the standardized examinees. Ten experienced and ten inexperienced examiners watched the video recording of the OSCE station on sigmoid diverticulitis. Experienced examiners had participated at least three times

Task 1			
Complete and independent taking of a patient's case history		Addressed if marked by X	Explanations
Pain	Description of pain	X	Do ask about the type, length, location and progression of the pain. Do not ask about earlier symptoms of this nature.
	Location of pain	X	
	Length of pain	X	
	Intensity of pain		
	Start of pain	X	
	Dependent on eating habits		

Figure 2: Example of an excerpt from a checklist defining the borderline examinee answers

Task 2		
Identify the five basic differential diagnoses!	Please not if the answer is given!	Remarks
Sigmoid diverticulitis		
Colon carcinoma		
Chronic inflammatory intestinal disease		
Unspecified colitis		
Renal colic		
Gynecological disease		

Figure 3: Example of an excerpt from a modified examiner checklist

or more in an OSCE and/or had more than five years of clinical experience. Inexperienced examiners were those who had served a maximum of two times as an OSCE examiner and/or had less than five years of clinical experience.

The original checklists from the surgical OSCE administered by the Medical Faculty of Heidelberg University were used to grade performance and required the assignment of one to five points for each item.

A briefing was held to impart general information on administering the test. The following instructions were given:

- The students perform a specific task which must be evaluated. No detailed information was given regarding the performance levels.
- The evaluation must be made based on what is contained in the checklists.
- Five points may only be assigned for a task if all items were accomplished without assistance.
- Three points may only be assigned for a task if all items were fully completed with the assistance of the examiner.
- One point may be given for a task if it was done incompletely despite the assistance of the examiner.
- Stopping and rewinding the video to view it again was not permitted.
- All four test situations must be viewed in sequence and without interruption.

The examiners were only informed after evaluating the videos that the students had been standardized to perform at a defined level.

Acceptance of standardized examinees

After evaluating all of the test situations, all of the examiners were surveyed to evaluate the acceptance of standardized examinees and their possible uses. The following was asked directly:

- Assessing the performance was easy for me.
- I would find it easier to assess in a real test situation.
- The assessment of the performance was difficult for me.
- The assessment of performance by good examinees was easy for me.
- The assessment of performance by poor examinees was easy for me.
- I find it makes sense to use standardized examinees to prepare inexperienced examiners.
- Training with video recordings (as opposed to training in a simulated OSCE) is sufficient to prepare examiners.
- Inexperienced examiners should be trained using standardized examinees before conducting real assessments.
- Experienced examiners should simulate test situations using standardized examinees.

- Targeted training of examiners using standardized examinees can make the OSCE objective.
- The performance of the standardized examinees was authentic.

The evaluation was done using a five-point Likert scale with 1=*completely disagree* to 5=*completely agree*.

Statistical analysis

Only a purely descriptive and qualitative analysis was carried out for the first part of the study due to the small cohorts and the individual approaches. Further statistical tests were not applied. The OSCE answer sheets were analyzed as to whether too many or too few answers had been given. Later, the study coordinator used the video recordings to analyze which difficulties arose when answering the questions. All of the quantitative analyses based on the OSCE checklists and the secondary video analysis were compiled and the percentages of deviations from the prescribed answers were calculated for all of the evaluations (see table 1).
For the second part of the study, the results of the comparison between experienced and inexperienced examiners are presented as mean values with standard deviation, if not otherwise indicated. The quantitative parameters were analyzed using the two-sided T-test. Categorical variables are given as absolute values. Statistical significance is assumed when the p-value is <0.05. Statistical analysis was carried out using IBM SPSS Statistics 25 software.

Results

First part of study: development of the standardized examinees

Verification of the standardization – descriptive analysis

An individual evaluation was carried out at the item level for each examinee. The percentage of deviations in the answers given from the expected number of responses was analyzed based on the standardization. In doing this, all of the evaluations, checklists from the OSCE, and the secondary quantitative video analysis by the study coordinator were compiled. The detailed results can be found in table 1. Only three examinees were analyzed for the *abdominal examination* checklist since one student was unable to participate in the OSCE for health reasons.
It became clear that especially students with a borderline performance had problems giving the answers correctly. The deviations were more distinct than in the case of the excellent students.
On the checklists covering *sigmoid diverticulitis and rectal carcinoma*, the difficulties were few for the excellent students: They gave a low percentage of too few answers. Larger deviations were seen for the borderline students. The largest deviation occurred for items 3 and 4. These items covered the determination of additional diagnostic and therapeutic measures.
The largest deviation was seen for the station on *abdominal examination* for item 4 by the students giving a borderline performance in the form of a high percentage of missing answers or incorrect performance of medical examination procedures. For this item, the examinees' examination of the liver was assessed. On this checklist, borderline students showed overall heterogeneous performances with too many and too few answers. Standardized examinees who gave an excellent performance, on the other hand, had a tendency to give too few answers or not to perform individual steps of the medical examination procedure.

Assessment of performance by the examiners

All of the examiners, with one exception, had the impression that these were real examinees and indicated they had perceived the standardized examinees as authentic. The excellent performance was recognized as such in all cases. The borderline performance was assessed as borderline six times; in all other run-throughs, however, it was deemed to be a poor performance.

Qualitative analysis via video analysis

The qualitative analysis of the OSCE videos revealed a series of aspects that had a limiting effect on the standardization. The examinees showed a certain tendency to recite the expected answers as if they were memorized lists. This applied more to the excellent examinees than to the borderline ones. Borderline examinees had difficulties staying in their roles particularly for complex items that required drawing on a diagnostic or therapeutic algorithm and not allowing the examiner to push them into giving more than the standardized answers. On the whole, the standardized examinees were able to do this well. At the same time, it was noticed that occasionally the role was over-exaggerated and, for instance, an intentionally hesitant behavior was acted out in a very pronounced manner. As a result, time became tight in individual test situations.
The conduct of the examiners also influenced the students' acting of their roles and the results of the standardization. As in real assessments, the examiners showed a tendency to repeat questions or give advice on doing individual tasks. Among other things, this increased the difficulty the students faced in consciously not giving answers. Based on the video analysis, it also became clear that one examiner did not award points for answers which were given or examination steps that were performed. In another situation, an examiner evaluated the response of a simulated patient as the answer given by the examinee.
Likewise, it was observed that simulated patients actively influence the assessment by asking their own questions and preventing the students from giving an answer.

Table 1: Presented here are the percentages of deviations (too many and too few answers given) from the number of expected answers per item calculated from the quantitative analysis of the OSCE as well as the video analysis differentiated in addition according to standardized performance level (↑ Percent of answers that were given as too many; ↓ Percent of answers that were given as too few). Only three students could be analyzed for the *abdominal examination* checklist because one student was unable to participate in the OSCE for health reasons.

Checklist & Performance level	Item 1		Item 2		Item 3		Item 4		Item 5	
	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓
Sigmoid diverticulitis										
Excellent	0	4.8%	0	11.1%	0	8.3%	0	8.3%	0	0
Excellent	0	2.4%	0	0	0	0	0	0	0	0
Borderline	27.7%	5.6%	0	0	0	50%	11.1%	0	0	0
Borderline	0	11.1%	0	0	0	33.3%	11.1%	44.4%	0	0
Rectal carcinoma										
Excellent	2.0%	15.7%	0	9.5%	0	0	0	8.3%	0	0
Excellent	5.9%	5.9%	0	4.8%	0	0	0	0	0	0
Borderline	45.9%	8.3%	8.3%	0	22.2%	11.1%	22.2%	0	0	0
Borderline	8.3%	12.5%	0	0	11.1%	0	22.2%	22.2%	0	0
Abdominal examination										
Excellent	0	24.4%	0	11.1%	0	0	0	0	0	0
Borderline	0	0	0	0	16.7%	16.7%	0	87.5%	16.7%	50%
Borderline	0	11.1%	0	16.7%	16.7%	0	0	62.5%	0	50%

Second part of the study: influence of examiner experience on performance assessment and acceptance of standardized examinees

Influence of examiner experience on performance assessment

Ten experienced and ten inexperienced examiners were included in the study, with one female and nine male examiners forming the experienced group and three female and seven male examiners forming the inexperienced group. All of the examiners assessed all of the standardized examinees in one test situation during the OSCE in the first half of the study. The details regarding examiner experience are presented in table 2.

In the assessment of the examinees with excellent performance there was no significant difference between experienced and inexperienced examiners (see table 3 and figure 4). In contrast, there was a significant difference between the two groups in their assessments of the borderline examinees. Inexperienced examiners tended to assess the performance lower than their experienced counterparts.

Both groups of examiners graded the social competence (item 5), despite identical standardization, lower for the borderline examinees than for the excellent ones (see table 3 and figure 5). This difference was statistically significant (4.80 vs. 4.13, $p<0.001$).

Acceptance of standardized examinees

Both groups of examiners perceived the standardized examinees to be authentic and viewed this new tool as an opportunity to make the OSCE even more objective. Both groups found it easier to assess the performance of good students than of borderline students, but still found no difficulties overall in assessing student performance.

The regular use of standardized examinees to train experienced examiners was favored more by the group of inexperienced examiners than by the experienced group (2.9 vs. 2.0). The detailed results are presented in figure 6.

Discussion

Detailed instructions on how to design, implement and ensure the quality of an OSCE and the resulting good, statistically measured results justify the use of this test format to assess and grade practical clinical skills at medical schools [9], [15], [16], [17], [18]. While OSCEs and OSPEs, to date, have been used primarily as internal university-specific assessments, the current discussion on including them in state medical examinations is making the need for widespread standardization very clear [19]. Despite established quality assurance measures, a variety of studies have been able to show that factors can potentially influence OSCE scores. Such studies often involve extensive staff resources, e.g. independent co-examiners, video analyses, etc. At the same time, it is impossible to eliminate individual influences stemming from examinees and examiners or to standardize these factors satisfactorily. Our aim was to develop a new tool for OSCE quality assurance by applying the concept of standardization to student performance, an approach that enables the identification of individual factors influencing the grading of student performance. Simultan-

Table 2: Examiner characteristics, age as mean values with standard deviation; all figures given as absolute values

Variable	Experienced Examiner (n=10)	Inexperienced Examiner (n=10)
Age	37.4 (3.89)	28.8 (2.35)
Years of clinical experience		
≤1 year	0	2
1-3 years	0	7
3-5 years	0	1
>5 years	10	0
Clinical experience		
Outpatient	10	7
Normal ward	10	9
Intensive care unit	10	7
Number of times as OSCE examiner	n=9	n=10
0 times	0	2
1-2 times	2	7
3-5 times	5	1
>5 times	2	0
OSCE experience as student	n=9	n=10
Yes	0	8
No	9	2

Table 3: Assessment results for excellent and borderline students according to examiner experience, presented as mean values (min.-max.)

	Experienced examiner (n=10)	Inexperienced examiner (n=10)	p-Value
Excellent students, Overall number of points	24.25 (20-25)	24.50 (23-25)	0.441
Item 1	4.80	4.85	0.687
Item 2	5.00	4.95	0.324
Item 3	4.95	5.00	0.324
Item 4	4.89	4.95	0.530
Item 5	4.85	4.75	0.442
Borderline students, Overall number of points	15.15 (12-19)	13.50 (7-17)	**0.035**
Item 1	2.50 (1.539)	1.60 (0.598)	**0.020**
Item 2	2.15 (0.617)	2.00 (0.918)	0.559
Item 3	2.10 (0.718)	1.55 (0.686)	**0.018**
Item 4	2.50 (0.761)	4.35 (0.813)	**<0.001**
Item 5	4.25 (0.851)	4.00 (0.973)	0.393

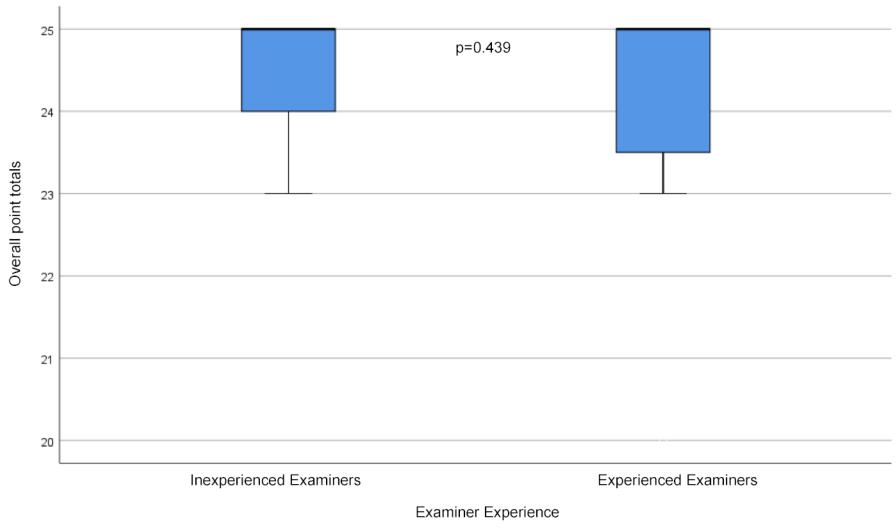

Figure 4: Overall point totals for the students with the excellent performance according the examiners' experience.

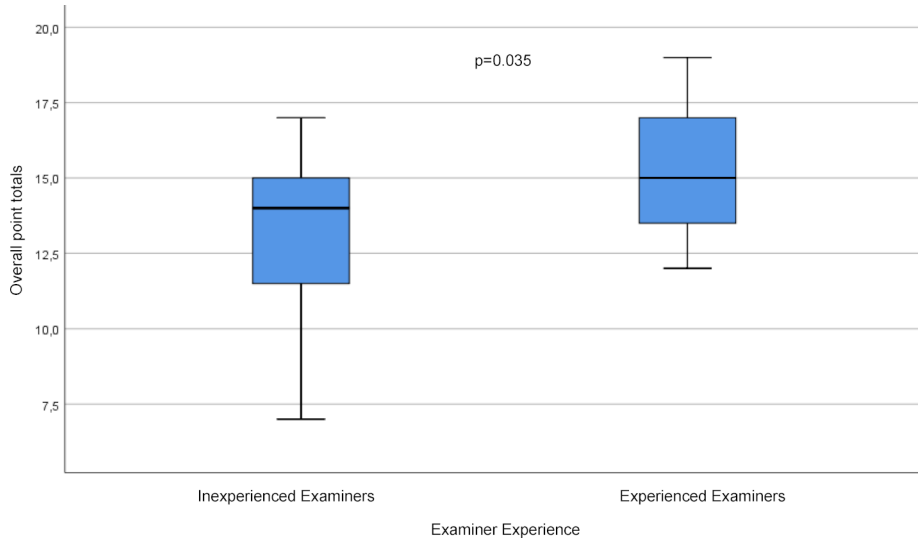

Figure 5: Overall point totals for the students with the borderline performance according to examiners' experience

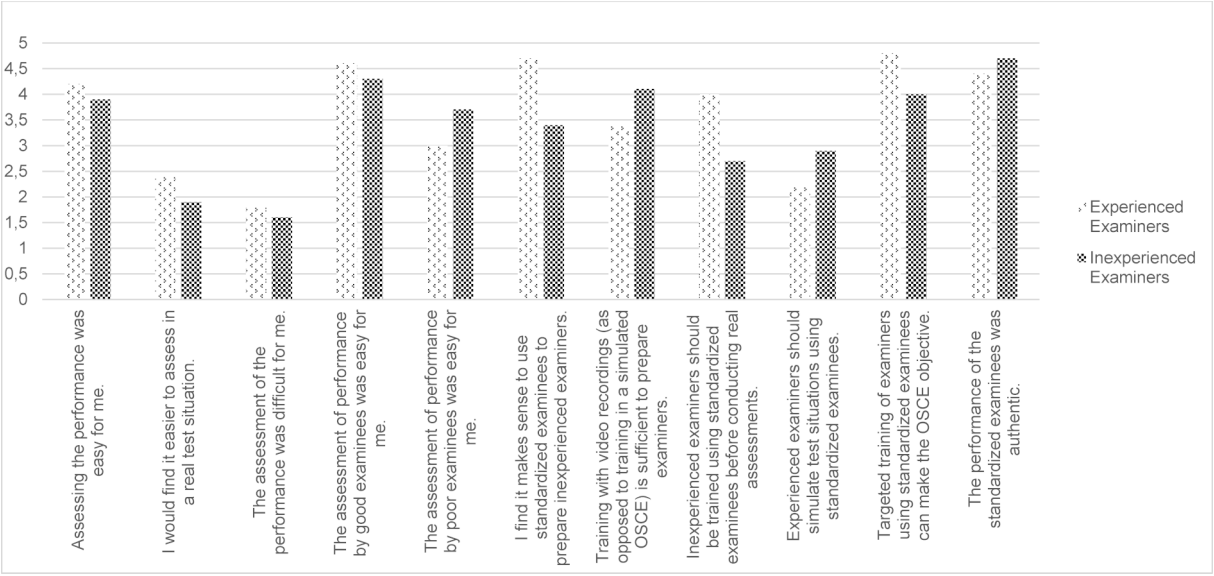

Figure 6: Results of the evaluation questionnaire regarding the standardized students according to experienced and inexperienced OSCE examiners, evaluation using a five-point Likert scale (1=completely disagree, 2=mostly disagree, 3=agree in part, 4=mostly agree, 5=completely agree)

eously, this new tool is also meant to serve as a strategy for training OSCE examiners in the future.

As part of verifying the standardized examinees, it was demonstrated that it is possible to successfully standardize students to meet a previously defined performance level.

The verification of the standardization revealed that deviations occurred in both groups of examinees. Excellent examinees tended more toward giving too few answers and had difficulties not appearing to recite previously memorized lists, while the borderline examinees gave both too few and too many answers. The deviations were overall more distinct for the borderline examinees indicating that the standardization for this performance level is more difficult to achieve.

The answers given by borderline examinees deviated in particular for items in which the description of a diagnostic or therapeutic algorithm was required (see table 1). This suggests that increased complexity of the task makes standardization more difficult. Similar observations were made regarding the complex examination procedures on the abdominal examination checklist where the borderline examinees also deviated from the expected procedural steps (see table 1). In addition to the purely content-based deviations, individual students tended to over-exaggerate their roles.

Both the content-based deviations by the standardized examinees and the different interpretations of the roles they played suggest that the process of standardization itself and specific training for the roles are essential. In the approach followed here, the students were trained using modified checklists on which, depending upon performance level, each possible answer was predefined and rehearsed, whether it was meant to be given or not. From these results it can be understood that the standardization should be trained in even more detail. As is

the case when training simulated patients [26], it appears sensible to define a larger role in which the performance level or the characteristic being assessed can be embedded. Since the examiners tended to repeat questions precisely for borderline examinees, the students must be very specially trained for such situations. In particular, attention must be paid to complex tasks and medical examination procedures. Based on the experiences described here, it is wise to let students repeatedly rehearse their roles for verification and to simulate different ways in which examiners intervene in the assessment process to practice conformity with the assigned roles on the part of the standardized examinees. Verifying standardization in a nearly real OSCE is also another option to check if standardization has been satisfactorily achieved. Video recording with subsequent analysis by the trainers and standardized examinees represents an additional training strategy.

An obvious disadvantage of this study is the low case numbers. The study involves one pilot project that is on par with a feasibility study. Future standardization of examinees should take place with more students and in a larger number of test situations than the selected number analyzed here.

In the second part of the study, the video recordings of the OSCE station addressing the *management of a patient* with sigmoid diverticulitis were used for both standardization levels. The extent to which examiner experience affected the evaluation of examinee performance was investigated. This station was used because the standardization for it was the best.

The results of this part of the study show that the two groups of examiners assessed the performance of borderline examinees differently. Inexperienced examiners graded the performance significantly lower and also applied a larger point range to do so. Basically, there are several conceivable explanations for this. Experienced examiners recognize the performance for what it is and correctly classify it as such. On the other hand, this observation could also indicate that experienced examiners do not use the full grading scale for recognizable performance levels and, as described by Iramaneerat, only apply a restricted range of points [19]. At the same time, this result could also be construed as indicating that inexperienced examiners are, under circumstances, less confident in classifying poor performances and thus rate them in a potentially exaggerated manner. A study by Yeates et al. demonstrated that different examiners focus on different aspects when assessing a performance [27]. The results here could therefore be a sign that with increasing clinical or assessment experience, the main focus for assigning points is selected unconsciously. It cannot be fully ruled out that all of the examiners here are not subject to a leniency error that is characterized by a general tendency to rate performances in an extreme manner as poorer or better than they actually are [13]. At the same time, it is possible that the effect described by Yeates et al. is present in that a borderline performance is rated especially poorly if it is observed directly after an excellent performance [7]. In the design selected here, the first and last performances in the video sequence were borderline performances, leaving only one instance where the constellation identified by Yeates et al. could have occurred.

The lower score assigned to social competence for borderline examinees (4.80 vs. 4.13, $p<0.001$), despite identical standardization and identical performance in the verification of standardization, leaves room to presume a halo effect for both examiner groups. The results of this study suggest that in terms of a halo effect, as described by Iramaneerat et al., the poorer content-based performance leads to a misperception of communication skills [21]. Experienced and inexperienced examiners were affected in equal measure by this, which points out that even having extensive experience as an OSCE assessor cannot negate this effect.

The detected differences in the assessment of borderline examinees depending on the examiner's experience suggest that this effect could potentially be decisive for passing or failing an OSCE station. The latter makes it clear that targeted examiner preparation is essential, especially if OSCEs are to be used in future state medical exams.

Another question that should be considered and explored in further studies is whether a difference exists in the grading behavior of experienced examiners depending on if they have experience as an OSCE assessor, or only have extensive clinical experience, or both. The experienced examiners in this study all had more than five years of clinical experience, but their experience as OSCE assessor varied between two and more than five times serving as OSCE examiners. This aspect was not pursued further since this study is a pilot project with a low case number.

In this study the use of videos to carry out such an analysis does not, by itself, present a novel approach. It is rather the standardized examinees who offer a possibility in the future to conduct very similar analyses in an OSCE with standardized examinees unconnected to video analyses. It is conceivable that standardized examinees could be included as a "quality standard" in an OSCE. The type of training for standardization must be explored and developed further to minimize deviations. Whether it is possible to standardize a student for several checklists still remains open.

Conclusions

Standardizing simulated examinees to meet defined performance levels represents a future possibility for directly analyzing influences on the grading behavior of OSCE examiners. Within the scope of high-stakes assessments, especially in regard to the future use of OSCEs in state medical exams, standardized examinees represent, alongside quality assurance, a potential tool to train and prepare OSCE examiners [19].

Competing interests

The authors declare that they have no competing interests.

References

1. Nikendei C, Kruppa E, Jünger J. Einsatz innovativer Lern- und Prüfungsmethoden an den Medizinische Fakultäten der Bundesrepublik Deutschland- eine aktuelle Bestandsaufnahme. Dtsch Med Wochenschr. 2009;134:731-732.
2. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J. 1975;22(1):447-451. DOI: 10.1136/bmj.1.5955.447
3. Schleicher I, Leitner K, Jünger J, Möltner A, Rüssler M, Bender B, Sterz J, Stibane T, König S, Frankenhauser S, Kreuder JG. Does quantity ensure quality? Standardized OSCE-stations for outcome-oriented evaluation of practical skills at medical faculties. Ann Anat. 2017;212:55-60. DOI: 10.1016/j.aanat.2017.03.006
4. Byrne A, Soskova T, Dawkins J, coombes L. A pilot study of marking accuracy and mental workload as measure of OSCE examiner performance. BMC Med Educ. 2016;16:191. DOI: 10.1186/s12909-016-0708-z
5. Wood TJ, Chan J, Humphrey-Murto S, Pugh D, Touchie C. The influence of first impressions on subsequent ratings within an OSCE station. Adv Health Sci Educ Theory Pract. 2017;22(4):969-983. DOI: 10.1007/s10459-016-9736-z
6. Fuller R, Homer M, Pell G, Hallam J. Managing extremes of assessor judgement within the OSCE. Med Teach. 2017;37(1):58-66. DOI: 10.1080/0142159X.2016.1230189
7. Yeates P, Cardell J, Byrne G, Eva KW. Relatively speaking: contrast effects influence assessors' scores and narrative feedback. Med Educ. 2015;49(9):909-919. DOI: 10.1111/medu.12777
8. Bartman I, Smee S, Roy M. A method of identifying extreme OSCE examiners. Clin Teach. 2013;10(1):27-31. DOI: 10.1111/j.1743-498X.2012.00607.x
9. Pell G, Fuller R, Homer M, Robert T. How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. Med Teach. 2010;32(10):802-811. DOI: 10.3109/0142159X.2010.507716
10. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81 Part I: A historical and theoretical perspective. Med Teach. 2013;35(9):e1437-1446. DOI: 10.3109/0142159X.2013.818634
11. Chesser A, Cameron H, Evans P, Gleland J, Boursicot K, Mires G. Sources of variation in performance on a shared OSCE station across four UK medical schools. Med Educ. 2009;43(6):526-532. DOI: 10.1111/j.1365-2923.2009.03370.x
12. Humphrey-Murto S, Touchi C, Wood TJ, Smee S. Does the gender of the standardised patient influence candidate performance in an objective structured clinical examination? Med Educ. 2009;43(6):521-525. DOI: 10.1111/j.1365-2923.2009.03336.x
13. Harasym PH, Woloschuk W, Cunning L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. Adv Health Sci Educ Theory Pract. 2008;13(5):617-632. DOI: 10.1007/s10459-007-9068-0
14. Turner JL, Dankosko ME. Objective structured clinical exams: A critical review. Fam Med. 2008;40(8):574-578.
15. Schultz JH, Nikendei C, Weyrich P, Möltner A, Fischer M R, Jünger J. Qualitätssicherung von Prüfungen am Beispiel des OSCE-Prüfungsformats: Erfahrungen der Medizinischen Fakultät der Universität Heidelberg. Z Evid Fortbild Qual Gesundhwes. 2008;102(10):668-672. DOI: 10.1016/j.zefq.2008.11.024
16. Barman A. Critiques on the objective structured clinical examination. Ann Acad Med Singapore. 2005;34(8):478-482.
17. Sloan DA, Donelly MB, Schwartz RW, Strodel WE. The Objective Structured Clinical Examination. The new gold standard for evaluating postgraduate clinical performance. Ann Sur.g 1995;222(6):735-742. DOI: 10.1097/00000658-199512000-00007
18. Mash B. Assessing clinical skill - standard setting in the objective structured clinical exam (OSCE). South Afr Fam Pract. 2007;49(3):5-7. DOI: 10.1080/20786204.2007.10873520
19. Jünger J. Kompetenzorientiert prüfen im Staatsexamen Medizin. Bundesgesundheitsbl. 2018;61:171-177. DOI: 10.1007/s00103-017-2668-9
20. Yeates P, O'Neill P, Mann K, Eva KW. 'You're certainly relatively competent': Assessor bias dur to recent experiences. Med Educ. 2013;47:910-922. DOI: 10.1111/medu.12254
21. Iramaneerat C, Yudkowsky R. Rater errors in a clinical skills assessmant of medical students. Eval Health Prof. 2007;30(3):266-283. DOI: 10.1177/0163278707304040
22. Schleicher I, Leitner K, Juenger H, Moeltner A, Ruesseler M, Bender B, Sterz J, Schuettler KF, Koenig S, Kreuder JG. Examiner effect on the objective structured clinicla exam - a study at five medical schools. BMC Med Educ. 2017;17:71. DOI: 10.1186/s12909-017-0908-1
23. Nikendei C, Kraus B, Lauber H, Schrauth M, Weyrich P, Zipfel S, Jünger J. An innovative model for teaching complex clinical procedures: Integration of standardised patients into ward round training for final year students. Med Teach. 2007;29(2-3):246-252. DOI: 10.1080/01421590701299264
24. Rethans JJ, Grosfeld FJ, Aper L, Reniers J, Westen JH, van Wijngaarden JJ, van Weel-Baumgarten EM. Six formats in simulated and standardized patients use, based on experiences of 13 undergraduate medical curricula in Belgium and the Netherlands. Med Teach. 2012;34(9):710-716. DOI: 10.3109/0142159X.2012.708466
25. Barrows HS. An Overview of the uses of standardized patients for teaching and evaluating clinical skills. Acad Med. 1993;68(6):443-451. DOI: 10.1097/00001888-199306000-00002
26. Schulz JH, Schönemann J, Lauber H, Nikendei C, Herzog W, Jünger J. Einsatz von Simulationspatienten im Kommunikations- und Interaktionstraining für Medizinerinnen und Mediziner (Medi-KIT): Bedarfsanalyse - Training - Perspektiven. Gruppendyn Organisationsberat. 2007;38(1):7-23. DOI: 10.1007/s11612-007-0002-y
27. Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently - Mechanisms that contribute to assessor differences in directly-observed performance assessments. Adv Helath Sci Educ Theory Paract. 2013;18(3):325-341. DOI: 10.1007/s10459-012-9372-1

Corresponding author:
Petra Zimmermann
Ludwig-Maximilians-Universität München, Klinikum der Universität, Klinik für Allgemein-, Viszeral- und Transplantationschirurgie, Marchionini Str. 15, D-81377 München, Germany, phone: +49 (0)89/4400-711239
petra.zimmermann@med.uni-muenchen.de

Please cite as
Zimmermann P, Kadmon M. Standardized examinees: development of a new tool to evaluate factors influencing OSCE scores and to train examiners. GMS J Med Educ. 2020;37(4):Doc40.
DOI: 10.3205/zma001333, URN: urn:nbn:de:0183-zma0013336

This article is freely available from
https://www.egms.de/en/journals/zma/2020-37/zma001333.shtml

Received: *2019-10-15*
Revised: *2020-02-23*
Accepted: *2020-04-27*
Published: *2020-06-15*

Standardisierte Prüflinge – Entwicklung eines neuen Instruments zur Beurteilung von Einflussfaktoren auf OSCE-Ergebnisse und zum Einsatz in der Prüferschulung

Petra Zimmermann[1]
Martina Kadmon[2]

1 Ludwig-Maximilians-Universität München, Klinikum der Universität, Klinik für Allgemein-, Viszeral- und Transplantationschirurgie, München, Deutschland

2 Universität Augsburg, Medizinische Fakultät, Gründungsdekanat, Augsburg, Deutschland

Zusammenfassung

Einleitung: Objective Structured Clinical Examinations (OSCE) sind als Format für klinisch-praktische Prüfungen an den meisten medizinischen Fakultäten etabliert und sollen in Zukunft auch in die humanmedizinischen Staatsprüfungen integriert werden. Einflüsse auf die Prüfungsergebnisse durch Prüferverhalten sind beschrieben. Fehlbeurteilungen der studentischen Leistungen resultieren beispielsweise durch systematische Nachsicht, durch Inkonsistenz in der Beurteilung, durch Halo-Effekte oder auch durch fehlende Differenzierung von Leistungen über die gesamte Bewertungsskala. Ziel der vorliegenden Arbeit war es ein Qualitätssicherungsinstrument zu entwickeln, das zukünftig die Überprüfung von Einflussfaktoren auf Bewertungen in einem realen OSCE ebenso wie eine gezielte Prüfer-Schulung ermöglicht.

Material, Methoden und Studierende: Zwölf Studierende der Medizinischen Fakultät Heidelberg wurden trainiert, eine definierte Leistung für jeweils eine chirurgische OSCE-Station zu erbringen. Es wurde ein Niveau für eine exzellente und eine Borderline-Leistung festgelegt und operationalisiert. Im ersten Teil der Studie wurde in einem Überprüfungs-OSCE die standardisierte Leistung dreimal hintereinander mit unterschiedlichen Prüfern/innen überprüft, bewertet und auf Video aufgenommen. Eine zusätzliche quantitative und qualitative Bewertung erfolgte durch die Studienleiterin anhand der Videoanalyse.

Im zweiten Teil der Studie wurden die Videoaufnahmen genutzt um die Akzeptanz für Standardisierte Prüflinge bei Prüfern/innen zu erheben und potentielle Einflüsse auf die Leistungsbewertung durch die Prüfererfahrung zu analysieren.

Ergebnisse: Im ersten Teil der Studie zeigten die Bewertungen im OSCE und die nachfolgende Videoanalyse, dass eine Standardisierung für definierte Leistungsniveaus an verschiedenen OSCE-Stationen grundsätzlich möglich ist. Einzelne Abweichungen von den erwarteten Antworten wurden beobachtet und traten vor allem mit zunehmender inhaltlicher Komplexität der OSCE-Station auf.

Im zweiten Studienteil bewerteten unerfahrene Prüfer/innen eine Borderline-Leistung signifikant schlechter als ihre erfahrenen Kolleg/innen (13,50 vs. 15,15, p=0,035). In der Bewertung der „Exzellenten Prüflinge“ zeigte sich kein Unterschied. Beide Prüfergruppen bewerteten das Item „Soziale Kompetenz“ – trotz identischer Standardisierung - bei Prüflingen mit einer Borderline-Leistung signifikant schlechter im Vergleich zu den „Exzellenten Prüflingen“ (4,13 vs. 4,80, p<0,001)

Schlussfolgerung: Die Standardisierung von Prüflingen für zuvor definierte Leistungsniveaus ist möglich, wodurch zukünftig ein neues Instrument sowohl zur Qualitätssicherung in OSCE-Prüfungen als auch zur Prüferschulung zur Verfügung steht. Eine detaillierte Vorbereitung der OSCE-Checklisten ebenso wie ein intensives Training mit den Prüflingen sind dabei unerlässlich.

Dieses neue Instrument gewinnt besondere Bedeutung, wenn standardisierte OSCE-Prüfungen in die medizinischen Staatsexamina integriert und somit als high-stakes Examen eingesetzt werden.

Schlüsselwörter: OSCE, OSPE, Prüferschulung, Qualitätssicherung, Standardisierte Prüflinge

Einleitung

Objective Structured Clinical Examinations (OSCEs) sind an den meisten medizinischen Fakultäten als Prüfungsform etabliert und eignen sich besonders zur Beurteilung klinisch-praktischer Fertigkeiten [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15]. In einem Leitlinienpapier der AMEE wurden verbindlich Standards und Messgrößen für die Qualitätssicherung von OSCEs definiert [9]. Für alle erforderlichen Prüfungen wird empfohlen, einen Blueprint sowohl für die Prüfungsinhalte als auch für eingesetzte Prüfungsformate zu erstellen. Grundlage für jeden OSCE sollte ebenfalls ein Blueprint sein, der die Prüfungsinhalte und zugeordnete Prüfungsstationen respektive Fachbereiche beinhaltet. Basierend auf dem Blueprint werden entsprechende Checklisten erstellt, diese in einem Review überprüft und die Leistungserwartungen anhand eines Standardsettings festgelegt. Durch eine ausreichende Anzahl an OSCE-Stationen, regelmäßige Standardsettings und Adaptationen der verwendeten Checklisten sowie regelmäßige Prüfereinweisungen wird eine gute Reliabilität und Interrater-Reliabilität erreicht. Teststatistische Auswertungen der Ergebnisse sollten herangezogen werden, um Probleme seitens der Checklisten oder der Prüfer/innen zu detektieren und durch regelmäßige Wiederholung des oben beschriebenen Prozesses zu minimieren [9], [15], [16], [17], [18].

Zahlreiche Untersuchungen analysieren potentielle Einflussfaktoren auf die Ergebnisse in einem OSCE. Diesen Einflussfaktoren kommt eine besondere Bedeutung zu, wenn das Prüfungsformat in High Stakes Prüfungen eingesetzt wird, wie es gerade in Deutschland für die medizinischen Staatsprüfungen in Diskussion ist [19]. Harasym und Kollegen konnten zeigen, dass Strenge oder Nachsichtigkeit seitens der Prüfer/innen zu einer systematischen zu schlechten oder zu guten Bewertung führen können [13]. Auch das Leistungsniveau eines Studierenden scheint die Reliabilität der Leistungsbewertung durch Prüfer/innen zu beeinflussen. Byrne et al. beschrieben, dass eine gute Studierenden-Leistung exakter bewertet wurde als eine Borderline-Leistung [4]. Yeates und Kollegen stellten in mehreren Untersuchungen fest, dass eine gute Leistung beispielsweise besser bewertet wird, wenn die zuvor bewertete Leistung schlecht war [7], [20]. Gleichzeitig wurde eine Borderline-Leistung schlechter bewertet, wenn der/die Prüfer/Prüferin zuvor eine gute Leistung beurteilt hatte. Darüber hinaus wurden Auswirkungen auf die Bewertung durch Halo-Effekte und fehlende Leistungsdifferenzierung über die gesamte Bewertungsskala beschrieben [21]. Schleicher und Kollegen konnten in einer Fakultäten-übergreifenden Untersuchung zeigen, dass studentische Leistungen unterschiedlich von lokalen und Referenz-Prüfern/innen bewertet wurden. Gleichzeitig zeigte sich ein Trend zu unterschiedlichen Bewertungen abhängig vom Geschlecht der Prüfer und der Prüflinge [22].

Alle bisherigen Untersuchungen zu potentiellen Einflussfaktoren und zur Qualitätssicherung des Prüfungsformats basieren auf Analysen der Ergebnisse aus live Beobachtungen oder Video-Analysen von OSCEs. Zwar basieren diese Analysen auf OSCEs, denen im Allgemeinen eine standardisierte Prüfereinweisung vorausging, potentielle Einflussgrößen seitens der Prüflinge unterliegen, aber keiner Standardisierung, sodass letztlich Prüfer-Eigenschaften nicht völlig isoliert beurteilt werden können.

Ein geeignetes Instrument, das es ermöglicht potentielle Einflussgrößen auf Seite des Prüflings zu simulieren, um so eine direkte Analyse der entstehenden Auswirkungen auf das Prüfer-Verhalten und die Ergebnisse zu ermöglichen, existiert bisher nicht. Gleichzeitig steht bislang kein geeignetes Instrument zur Verfügung, um Prüfer/innen im Hinblick auf potentielle Einschränkungen in der Reliabilität der Bewertung von Leistungen in einem OSCE gezielt zu schulen.

Simulationspatienten stellen mittlerweile einen integralen Bestandteil der medizinischen Ausbildung und auch medizinischer Prüfungen dar. Sie bieten die Möglichkeit Gesprächs- und Untersuchungssituationen in einem geschützten Rahmen zu üben und können eine Rolle immer wieder in standardisierter Weise spielen. Gleichzeitig besteht hierdurch die Möglichkeit einzelne Parameter, z.B. die Reaktion des Simulationspatienten oder das Ausmaß der Erkrankung, zu variieren um so unterschiedliche Situationen für den Studierenden zu simulieren [23], [24], [25].

Basierend auf dem Konzept der Simulationspatienten war es unser Ziel, dieses Konzept der Standardisierung auf die studentische Leistung in einem OSCE zu übertragen. Im ersten Studienteil der vorliegenden Arbeit wird überprüft, ob es möglich ist, Studierende zu trainieren eine definierte Leistung wiederholt in einem OSCE zu erbringen. Im zweiten Studienteil wird anhand der generierten Videosequenzen aus dem ersten Studienteil der Einfluss der Prüfererfahrung auf die Leistungsbewertung analysiert und die prinzipielle Akzeptanz für Standardisierte Prüflinge unter Prüfern/innen evaluiert.

Hierdurch konnte ein neues Instrument zur Qualitätssicherung in einem OSCE etabliert werden, das gleichzeitig ermöglicht, einzelne Einflussfaktoren auf die Bewertung zu identifizieren und Prüfer/innen zukünftig gezielt zu schulen.

Material, Methoden & Studierende

Zwölf Studierende wurden für eine standardisierte Leistung an drei verschiedenen Stationen der chirurgischen OSCE-Prüfung an der Medizinischen Fakultät Heidelberg trainiert. Pro Station wurden jeweils 2 Studierende für eine hervorragende (Exzellente Leistung) und zwei Studie-

rende für eine grenzwertige Leistung (Borderline-Leistung) standardisiert, jeweils eine weibliche Studierende und ein männlicher Studierender pro Leistungsniveau. Ein Studierender, der für eine exzellente Leistung für die OSCE-Station „Abdominelle Untersuchung“ vorbereitet war, konnte krankheitsbedingt kurzfristig nicht an der Studie teilnehmen.
Als exzellente Leistung wurde das Erreichen der Höchstpunktzahl mit einem maximalen Abzug von 2 Punkten definiert, als Borderline-Leistung das Erreichen der minimal erwarteten Punktzahl für das Bestehen der jeweiligen Checkliste (Minimalkompetenz) ± 1 Punkt.
Die Summe aller Minimalkompetenzen innerhalb des Heidelberger chirurgischen OSCE, stellt die Bestehensgrenze für den Gesamt-OSCE dar.
Abbildung 1 stellt schematisch das Studiendesign dar, Abbildung 1A beschreibt den ersten und Abbildung 1B den zweiten Studienteil.

OSCE Checklisten

Es wurden drei bereits im chirurgischen OSCE gut etablierte und mehrfach in internen Reviews überprüfte Checklisten ausgewählt. Die Checklisten bezogen sich auf folgende OSCE-Stationen:

- *Management eines Patienten mit Sigmadivertikulitis*
- *Management eines Patienten mit V.a. Rektumkarzinom*
- *Abdominelle Untersuchung*

Alle Checklisten sind auf eine Minimalpunktzahl von 0 und auf eine Maximalpunktzahl von 25 Punkten ausgelegt. Jede Checkliste besteht aus 5 Teilaufgaben (Items), die jeweils mit maximal 5 Punkten bewertet werden können. Jedes Item umfasst unterschiedlich viele geforderte Antworten.
Die Minimalkompetenz bezeichnet die Punktzahl, die zum Bestehen der einzelnen Checkliste erreicht werden muss. Sie ist als minimale Erwartung an der jeweiligen Station auf der Basis der vorliegenden Checkliste definiert. Sie wird regelmäßig überprüft und im internen Standardsetting festgelegt. Die Minimalkompetenzen für die hier verwendeten Checklisten liegen bei 17 Punkten.
Die maximale Prüfungsdauer pro Checkliste beträgt 9 Minuten, die Wechselzeit zur nächsten Station eine Minute. Auf den Checklisten sind die übergeordneten Bewertungskategorien (z.B. Anamneseerhebung, klinische Untersuchung, etc.) und zugeordnete Einzelitems zur Punktevergabe ausgewiesen:

- 5 Punkte: sämtliche Leistungen ohne Hilfe erbracht
- 3 Punkte: sämtliche Leistungen mit Hilfe des Prüfers vollständig erbracht
- 1 Punkt: Leistungen mit Hilfe des Prüfers unvollständig erbracht

Für jede Bewertungskategorie ist angegeben, ob eine Punktevergabe global für den Gesamteindruck oder auf der Basis von Antworten auf die Einzelitems erfolgen soll. Jede Checkliste enthält eine kurze Fallvignette sowie pro Einzelitem eine Aufgabenstellung und die erwarteten Antworten. Mögliche Zwischenfragen durch die Prüfer/innen sind nicht vordefiniert.
Die Checklisten zur *Sigmadivertikulitis* und zum *Rektumkarzinom* beinhalten ein Anamnesegespräch mit einem Standardpatienten (Item 1), die Ableitung von Differentialdiagnosen aus den anamnestischen Details (Item 2), die Entscheidung, welche geeigneten diagnostischen Maßnahmen in der konkreten Situation eingeleitet werden sollen (Item 3) sowie bei der *Sigmadivertikulitis* die Beschreibung eines CT-Ausschnitts zu dem Patientenfall. Item 4 umfasst bei beiden Checklisten wieder die Interaktion mit dem Standardpatienten zum weiteren diagnostischen/therapeutischen Vorgehen. Item 5 beurteilt die soziale Kompetenz. Dabei wird unter anderem beurteilt in wie weit der Studierende sich dem Patienten adäquat vorgestellt hat, sich gegenüber dem Patienten verhält, z.B. ob Blickkontakt gehalten werden kann.
Die Checkliste *Abdominelle Untersuchung* umfasst sequentiell eine abdominelle Untersuchung bei einem Patienten mit rechtsseitigen Unterbauchschmerzen (Item 1), Überprüfung der Peritonitiszeichen (Item 2), die Erläuterung zur Durchführung und Befundung einer digital-rektalen Untersuchung (Item 3), die Untersuchung der Leber (Item 4) und die Untersuchung der Milz (Item 5).

Modifikation der OSCE Checklisten

Zur Durchführung der Standardisierung der Leistung der Standardisierten Prüflunge sowie zur Überprüfung, ob diese Leistung mehrfach wiederholt werden kann, wurden aus den für den chirurgischen OSCE vorliegenden Checklisten jeweils 2 neue Versionen generiert.

Checklisten zur Standardisierung der Prüflinge

Zur Standardisierung der Prüflinge wurden alle Checklisten detailliert operationalisiert. Bezogen auf die beiden festgelegten Leistungsniveaus wurde für jede mögliche Antwort eines Items der Checkliste definiert, ob die Prüflinge diese nennen sollten oder nicht. Gleichzeitig wurde in einem weiteren Feld für jedes Item ausgeführt, wie der Prüfling sich bei der jeweiligen Frage verhalten sollte, z.B. nur zögerliche Antwort, oder nur auf Nachfrage (siehe Abbildung 2).

Checklisten zur Leistungsüberprüfung

Zur Beurteilung der erbrachten Leistung durch die Prüfer/innen wurde der Bewertungsteil die OSCE Checklisten so modifiziert, dass die Prüfer/innen für jede mögliche Antwort in jedem Aufgabenbereich vermerken konnten, ob die Antwort gegeben wurde oder nicht (siehe Abbildung 3). Es wurde bewusst darauf verzichtet, die Standardisierung über die Beurteilung des Leistungsniveaus mit einem globalen Punktwert für jedes Item, analog zu einem echten OSCE, durchzuführen, um potentielle systematische Einschätzungsunterschiede Seitens der Prüfer/innen zu eliminieren. Am Ende der Checkliste wurde ein Abschnitt eingefügt, in dem die Prüfer/innen das erbrachte Leis-

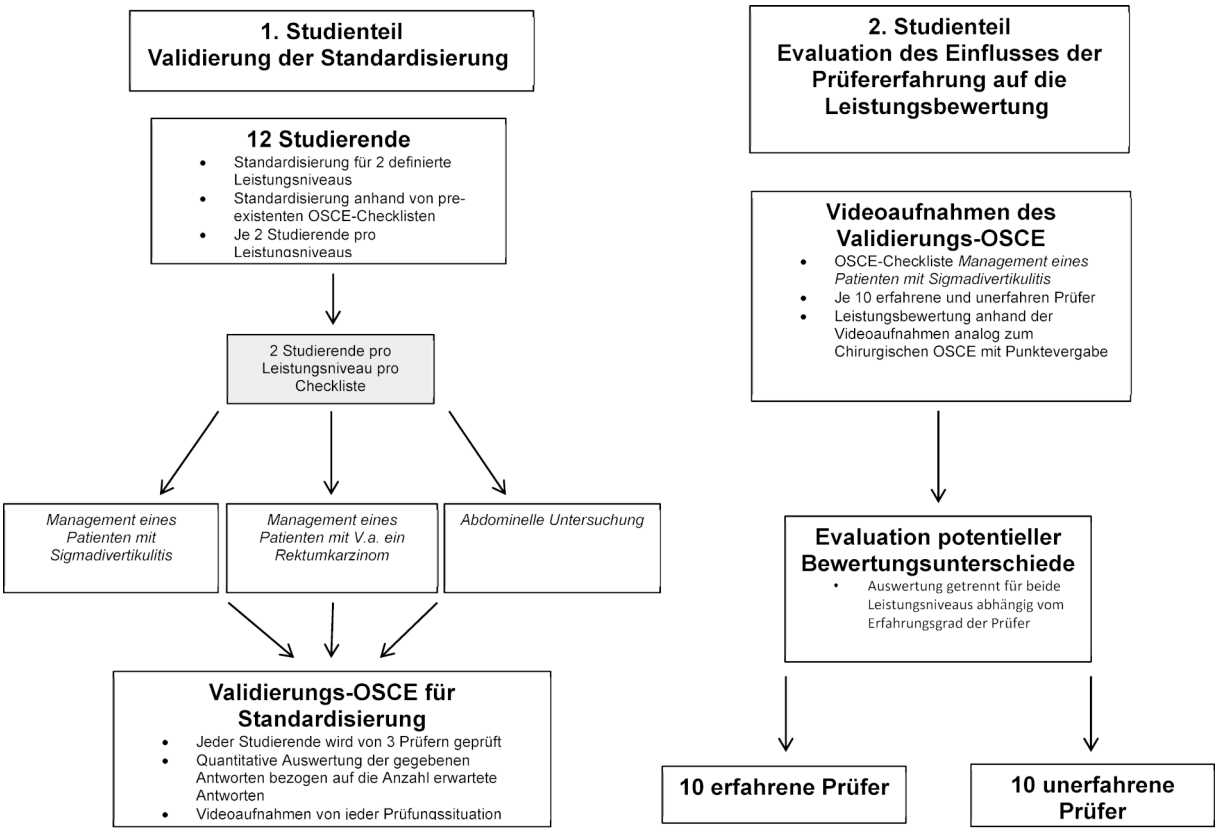

Abbildung 1: Schematische Darstellung des Studienablaufs mit 1. und 2. Studienteil

Aufgabe 1			
Anamneseerhebung vollständig und eigenständig		Punkte nennen, wenn Kreuz	Erläuterungen
Schmerzen	Schmerzcharakter	X	Fragen Sie nach Art, Dauer, Lokalisation und Verlauf der Schmerzen Fragen Sie nicht nach früheren Beschwerden dieser Art
	Schmerzlokalisation	X	
	Schmerzdauer	X	
	Schmerzintensität		
	Schmerzbeginn	X	
	Nahrungsabhängigkeit		

Abbildung 2: Exemplarischer Auszug aus einer Checkliste mit Definitionen für die Antworten der Prüflinge mit einer Borderline-Leistung

Aufgabe 2		
Nennen Sie die 5 wesentlichen Differentialdiagnosen!	Bitte kreuzen, wenn Punkt genannt!	Ggf. Erläuterungen
Sigmadivertikulitis		
Colon-Ca		
Chronisch-entzündliche Darmerkrankungen		
Unspezifische Colitis		
Harnleiterkolik		
Gynäkologische Erkrankungen		

Abbildung 3: Exemplarischer Auszug aus einer modifizierten Prüfer-Checkliste

tungsniveau anhand einer globalen Bewertungsskala (schlecht, mittelmäßig, sehr gut) ebenso wie die Authentizität beurteilen sollten. Bei letzterem sollten die Prüfer/innen beurteilen, in wieweit sie Zweifel daran hatten, einen echten Prüfungskandidaten vor sich zu haben.
Die Prüfer/innen erhielten die standardisierte Prüfereinweisung für den Chirurgischen OSCE. Sie wurden jedoch angewiesen keine Punkte für die einzelnen Items zu vergeben, sondern für jede mögliche Antwort anzukreuzen, ob diese gegeben wurde oder nicht. Den Prüfern/innen wurde erst nach dem OSCE mitgeteilt, dass eine Standardisierung der Studierenden-Leistung vorgenommen worden war.

Standardisierte Studierende

Alle 12 Studierenden hatten den Chirurgischen Block und damit den chirurgischen OSCE bereits absolviert. Der Chirurgische Block erstreckt sich über ein Semester und beinhaltete die Fächer Viszeralchirurgie, Gefäß-, Thorax- und Herzchirurgie, Urologie, Orthopädie & Unfallchirurgie, Hand- & Plastische Chirurgie sowie Anästhesie & Notfallmedizin. Vorlesungen und Seminare in Pathologie und Radiologie sind in den einzelnen Fachdisziplinen integriert.
Den Studierenden wurden die jeweiligen Checklisten zum Training ausgehändigt. Mit jedem Studierenden wurden die Rolle und die erwarteten Antworten anhand der modifizierten Checkliste detailliert durchgesprochen. Nach 2 Wochen Zeit zum Lernen der Checkliste und Rolle wurde die Prüfungssituation zwischen den Studierenden und der Studienleiterin simuliert und Korrekturen umgesetzt. Dabei wurden zunächst allgemeine Schwierigkeiten und Aspekte besprochen, nachfolgend die Prüfungssituation in Echtzeit simuliert und abschließend nochmals ein Feedback zu erforderlichen Anpassungen gegeben.

1. Studienteil

Durchführung der Standardisierung

Im ersten Studienteil (siehe Abbildung 1, links A) erfolgte die Durchführung der Standardisierung in einem simulierten OSCE, der entsprechend realer Prüfungsbedingungen (Zeit, Wechselzeiten, etc.) durchgeführt wurde. Die Standardisierten Prüflinge spielten ihre Rolle dreimal mit drei unterschiedlichen Prüfern/Prüferinnen (ein Prüfer und zwei Prüferinnen) und wurden dabei auf Video aufgenommen. Zusätzlich wurden in einem zweiten Schritt alle Videoaufnahmen durch die Studienleiterin mit Hilfe der modifizierten Prüferchecklisten sowohl quantitativ als auch qualitativ ausgewertet, sodass für jeden Studierenden 6 Auswertungen vorlagen.
Bei der quantitativen Auswertung wurden basierend auf den zuvor festgelegten Antworten, die gegeben werden sollten, die Abweichungen gezählt. Dabei wurden sowohl zu viel als auch zu wenig gegebene Antworten bezogen auf die korrekt erwartete Anzahl Antworten berücksichtigt. Nachfolgend wurden die durchschnittlichen prozentualen Abweichungen für alle OSCE-Durchläufe (3 Prüfungssituationen) sowie für die quantitative Auswertung aus der nachfolgenden Video-Auswertung berechnet.
Bei der qualitativen Auswertung wurde zunächst der Globaleindruck bewertet: Der Prüfling wirkt authentisch ja/nein und agiert im Rahmen seiner Rolle. Zusätzliche wurden folgende Aspekte beurteilt:

- Verhalten des Prüflings beim Geben der Antworten (wirkt sicher, unsicher, neigt dazu Listen widerzugeben)
- Reaktion des Prüflings auf Verhalten/Fragen des/der Prüfers/Prüferin (bleibt in der Rolle, weicht von erwarteten Antworten ab, lässt sich zu Antworten drängen)
- Reaktion des Prüflings auf Verhalten/Fragen des/der Standardpatienten/in (bleibt in der Rolle, weicht von erwarteten Antworten ab, lässt sich zu Antworten drängen)
- Verhalten der Prüfer/innen
- Verhalten der Standardpatienten/innen

Die Studienleiterin selbst war für die Organisation des Chirurgischen Blocks mit verantwortlich und hat mehr als 20 Mal in einem OSCE des Chirurgischen Blocks geprüft. Zusätzlich hat sie Erfahrung in der Erstellung von OSCE-Checklisten und Prüfungsfragen. Diese Studie erfolgte im Rahmen ihrer Masterthese für den Master of Medical Education, Deutschland (MME-D).

2. Studienteil

Analyse des Einflusses der Prüfererfahrung auf die Leistungsbewertung

Im zweiten Studienteil (siehe Abbildung 1, rechts B) wurde mit Hilfe der Videoaufnahmen der Einfluss der Prüfererfahrung auf die Leistungsbewertung sowie die Akzeptanz für Standardisierte Prüflinge auf Prüferseite untersucht. Je 10 erfahrenen und unerfahrenen Prüfern und Prüferinnen wurden die Videoaufnahmen der OSCE-Station *Sigmadivertikulitis* gezeigt. Erfahrene Prüfer bzw. Prüferinnen hatten mindestens 3 Mal oder mehr als Prüfer in einem OSCE teilgenommen und/oder mehr als 5 Jahre klinische Erfahrung. Unerfahrene Prüfer und Prüferinnen waren diejenigen, die maximal 2 Einsätze als OSCE-Prüfer und/oder weniger als 5 Jahre klinische Erfahrung hatten. Zur Leistungsbeurteilung wurde die Originalchecklisten aus dem chirurgischen OSCE der Medizinischen Fakultät Heidelberg, die eine Punktevergabe (1-5) pro Item erfordern, verwendet.
Es erfolgte eine Einweisung mit allgemeinen Informationen zum Ablauf der Prüfung. Im Einzelnen wurde darauf hingewiesen, dass

- die Studierenden eine bestimmte Leistung erbringen, die beurteilt werden soll. Es wurden keine Detailinformationen zu den Leistungsniveaus genannt.
- die Beurteilung basierend auf dem Inhalt der Checkliste erfolgen muss.
- 5 Punkte für eine Aufgabe nur vergeben werden dürfen, wenn sämtliche Leistungen ohne Hilfe erbracht wurden.

- 3 Punkte für eine Aufgabe nur dann vergeben werden dürfen, wenn sämtliche Leistungen mit Hilfe des Prüfers vollständig erbracht wurden.
- 1 Punkt für eine Aufgabe vergeben werden kann, wenn die Leistung mit Hilfe des Prüfers unvollständig erbracht wurde.
- das Anhalten und erneute Abspielen des Videos nicht erlaubt sind.
- alle 4 Prüfungssituationen hintereinander angeschaut werden müssen ohne Unterbrechung.

Den Prüfern und Prüferinnen wurden erst nach Beurteilung aller Videos mitgeteilt, dass die Studierenden standardisiert waren eine definierte Leistung zu erbringen.

Akzeptanz für Standardisierte Prüflinge

Nach Beurteilung aller Prüfungssituationen erhielten alle Prüfer/innen einen Fragebogen zur Evaluation der Akzeptanz für Standardisierte Prüflinge und ihrer Einsatzmöglichkeiten. Konkret wurden folgenden Punkten abgefragt:

- Die Einschätzung der Leistung ist mir leicht gefallen.
- In einer realen Prüfungssituation fände ich die Einschätzung der Leistung leichter.
- Die Einschätzung der Leistung ist mir schwer gefallen.
- Bei guten Prüflingen ist mir die Einschätzung der Leistung leicht gefallen.
- Bei schlechten Prüflingen ist mir die Einschätzung der Leistung leicht gefallen.
- Ich halte den Einsatz von Standardisierten Prüflingen als Vorbereitung für unerfahrene Prüfer für sinnvoll.
- Ein Training mit Video-Aufnahmen (im Gegensatz zum Training in einem simulierten OSCE) ist ausreichend zur Prüfer-Vorbereitung.
- Unerfahrene Prüfer sollten vor dem Einsatz in realen Prüfungen an Standardisierten Prüflingen trainiert werden.
- Erfahrene Prüfer sollten an Standardisierten Prüflingen Prüfungssituationen simulieren.
- Ein gezieltes Prüfer-Training an Standardisierten Prüflingen kann den OSCE objektivieren.
- Die Leistung der Standardisierten Prüflinge war authentisch.

Die Bewertung erfolgte mit einer 5-Punkte Likert-Skala von 1=*trifft gar nicht zu* bis 5=*trifft voll zu*.

Statistische Auswertung

Aufgrund der kleinen Kohorte sowie des individuellen Ansatzes erfolgte im ersten Studienteil eine rein deskriptive und qualitative Auswertung. Auf weitere statistische Tests wurde verzichtet. Die Antwortbögen des OSCEs wurden im Hinblick auf zu viel bzw. zu wenig gegebene Antworten ausgewertet. Nachträglich wurde per Videoanalyse durch die Studienleiterin beurteilt, welche Schwierigkeiten bei der Beantwortung der Fragen auftraten. Alle quantitativen Auswertungen basierend auf den OSCE-Checklisten sowie der sekundären Videoauswertung wurden zusammengefasst und die prozentuale Abweichung von den erwarteten Antworten für alle Beurteilungen berechnet (siehe Tabelle 1).
Im zweiten Studienteil wurden die Ergebnisse des Vergleichs zwischen erfahrenen und unerfahrenen Prüfern/innen in Mittelwerten mit Standardabweichung dargestellt, falls nicht anders angegeben. Die quantitativen Parameter wurden mit Hilfe des zweiseitigen t-Test analysiert. Kategorische Variablen werden als Absolutwerte angegeben. $P<0,05$ wurde als statistisch signifikant angenommen. Zur statistischen Berechnung wurde die IBM SPSS Statistics 25 Software verwendet.

Ergebnisse

Erster Studienteil – Entwicklung der Standardisierten Prüflinge

Überprüfung der Standardisierung – deskriptive Auswertung

Für jeden Prüfling erfolgte eine Einzelauswertung auf Itemniveau. Analysiert wurde die prozentuale Abweichung der gegebenen Antworten von der erwarteten Anzahl der Antworten auf der Basis der Standardisierung. Dabei wurden alle Auswertungen, Checklisten aus dem OSCE sowie die sekundäre quantitative Auswertung der Videoanalyse durch die Studienleiterin, zusammengefasst. Die detaillierten Ergebnisse sind in Tabelle 1 dargestellt. Für die Checkliste *Abdominelle Untersuch*ung wurden nur drei Prüflinge ausgewertet werden, da ein Studierender krankheitsbedingt nicht am OSCE teilnehmen konnte.
Es kristallisierte sich heraus, dass insbesondere die Studierenden mit einer Borderline-Leistung Schwierigkeiten hatten, die Antworten korrekt zu geben. Die Abweichungen waren deutlicher als bei den exzellenten Studierenden.
Bei den Checklisten *Sigmadivertikulitis* und *Rektumkarzinom* waren die Schwierigkeiten für die exzellenten Studierenden gering, sie nannten in einem geringen Prozentsatz zu wenig Antworten. Bei den Borderline Studierenden fielen größere Abweichungen auf. Die größte Abweichung trat bei den Items 3 und 4 auf. Diese Items umfassen die Festlegung des weiteren diagnostischen bzw. therapeutischen Vorgehens.
Die größte Abweichung zeigte sich bei der Station *Abdominelle Untersuchung* bezüglich Item 4 für die Studierenden mit einer Borderline-Leistung im Sinne eines hohen Anteils an fehlenden Antworten bzw. nicht korrekt durchgeführten Untersuchungsabläufen. Bei diesem Item wird die Untersuchung der Leber geprüft. Borderline Studierenden zeigten ansonsten bei dieser Checkliste insgesamt heterogene Leistungen mit zu vielen und zu wenig Antworten. Standardisierte Prüflinge mit einer exzellenten Leistung hatten wiederum die Tendenz zu wenig Antworten zu nennen bzw. bei den Untersuchungsabläufen einzelne Punkte nicht durchzuführen.

Tabelle 1: Dargestellt ist die prozentuale Abweichung (zu viel und zu wenig genannte Antworten) von der Anzahl erwarteter Antworten pro Item berechnet aus der quantitativen Auswertung des OSCE und der Videoanalyse unterschieden zusätzlich nach standardisiertem Leistungsniveau (↑ Anteil Antworten, der zu viel genannt wurde; ↓ Anteil Antworten, der zu wenig genannt wurde). Für die Checkliste Abdominelle Untersuchung konnten nur drei Studierende ausgewertet werden, da ein Studierender krankheitsbedingt nicht am OSCE teilnehmen konnte.

Checkliste & Leistungsniveau	Item 1		Item 2		Item 3		Item 4		Item 5	
	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓
Sigma-divertikulitis										
Exzellent	0	4,8%	0	11,1%	0	8,3%	0	8,3%	0	0
Exzellent	0	2,4%	0	0	0	0	0	0	0	0
Borderline	27,7%	5,6%	0	0	0	50%	11,1%	0	0	0
Borderline	0	11,1%	0	0	0	33,3%	11,1%	44,4%	0	0
Rektum-Karzinom										
Exzellent	2,0%	15,7%	0	9,5%	0	0	0	8,3%	0	0
Exzellent	5,9%	5,9%	0	4,8%	0	0	0	0	0	0
Borderline	45,9%	8,3%	8,3%	0	22,2%	11,1%	22,2%	0	0	0
Borderline	8,3%	12,5%	0	0	11,1%	0	22,2%	22,2%	0	0
Abdominelle Untersuchung										
Exzellent	0	24,4%	0	11,1%	0	0	0	0	0	0
Borderline	0	0	0	0	16,7%	16,7%	0	87,5%	16,7%	50%
Borderline	0	11,1%	0	16,7%	16,7%	0	0	62,5%	0	50%

Leistungseinschätzung durch die Prüfer/innen

Alle Prüfer und Prüferinnen hatten, mit einer Ausnahme, den Eindruck, dass es sich um reale Prüfungskandidaten handelte und gaben an, die standardisierten Studierenden als authentisch wahrgenommen zu haben.
Die exzellente Leistung wurde in allen Fällen als solche erkannt. Die Borderline-Leistung wurde 6 Mal als solche eingeschätzt, in allen anderen Durchläufen aber als schlechte Leistung wahrgenommen.

Qualitative Auswertung durch Video-Analyse

Die qualitative Auswertung der Prüfungsvideos ergab eine Reihe von Aspekten, die sich einschränkend auf die Standardisierung auswirkten. Die Prüflinge zeigten eine gewisse Tendenz, erwartete Antworten als auswendig gelernte Liste wiederzugeben. Diese betraf die exzellenten mehr als die Borderline-Prüflinge. Borderline-Prüflinge hatten besonders bei komplexen Items, die die Ableitung eines diagnostischen oder therapeutischen Algorithmus forderten, Schwierigkeiten in der Rolle zu bleiben und sich nicht durch den Prüfer bzw. durch die Prüferin zu mehr als den standardisierten Antworten drängen zu lassen. Insgesamt gelang dies den Standardisierten Prüflingen allerdings gut. Gleichzeitig fiel auf, dass gelegentlich die Rolle „überinterpretiert“ wurde und ein angedachtes zögerliches Verhalten zum Beispiel sehr ausgeprägt gespielt wurde. Dadurch wurde in einzelnen Prüfungssituationen die Zeit knapp.

Auch das Verhalten der Prüfer/innen beeinflusste die Wiedergabe der Studierenden-Rolle bzw. die Ergebnisse der Standardisierung. Wie in realen Prüfungen zeigten Prüfer/innen die Tendenz, z.B. nochmals nachzufragen oder kleinere Hinweise bei Einzelaufgaben zu geben. Dadurch erhöhte sich unter anderem die Schwierigkeit für die Studierenden, Antworten bewusst nicht zu geben. Anhand der Videoanalyse wurde außerdem deutlich, dass seitens eines Prüfers gegebene Antworten bzw. durchgeführte Untersuchungsschritte nicht gewertet wurden, obwohl sie erfolgt waren. In einer anderen Situation wertete ein Prüfer die Antwort eines Simulationspatienten als gegebene Antwort für den Prüfling.
Ebenso zeigte sich ein Einfluss durch die Simulationspatienten, die durch Zwischenfragen aktiv in die Prüfung eingriffen und dadurch ein Antworten des Studierenden verhinderten.

Zweiter Studienteil – Einfluss der Prüfererfahrung auf die Leistungsbewertung und Akzeptanz für Standardisierte Prüflinge

Einfluss der Prüfererfahrung auf die Leistungsbewertung

Zehn erfahrene und 10 unerfahrene Prüfer und Prüferinnen wurden in die Studie eingeschlossen, davon 1 weiblicher und 9 männliche Prüfer in der Gruppe der erfahrenen und 3 weibliche sowie 7 männliche Prüfer in der Gruppe der unerfahrenen Prüfer. Alle Prüfer/innen beur-

Tabelle 2: Prüfer-Charakteristika, Alter als Mittelwerte mit Standardabweichung, alle übrigen Angaben als Absolut-Werte

Variable	Erfahrene Prüfer (N=10)	Unerfahrene Prüfer (N=10)
Alter	37,4 (3,89)	28,8 (2,35)
Dauer der klinischen Erfahrung		
≤1 Jahr	0	2
1-3 Jahre	0	7
3-5 Jahre	0	1
>5 Jahre	10	0
Klinische Erfahrung		
Ambulanz	10	7
Normalstation	10	9
Intensivstation	10	7
Anzahl Einsätze als OSCE-Prüfer	n=9	n = 10
0 Einsätze	0	2
1-2 Einsätze	2	7
3-5 Einsätze	5	1
>5 Einsätze	2	0
OSCE-Erfahrung als Studierender	n=9	n=10
Ja	0	8
nein	9	2

teilten alle Standardisierten Prüflinge in einer Prüfungssituation aus dem OSCE des 1. Studienteils. Details zum Erfahrungsgrad der Prüfer/innen sind in Tabelle 2 angegeben.

In der Beurteilung der Prüflinge mit exzellenter Leistung zeigte sich kein wesentlicher Unterschied zwischen erfahrenen und unerfahrenen Prüfern/innen (siehe Tabelle 3, siehe Abbildung 4). Im Gegensatz dazu bestand ein signifikanter Unterschied in der Beurteilung der Borderline-Prüflinge zwischen den Prüfergruppen (siehe Tabelle 3, siehe Abbildung 5). Unerfahrene Prüfer und Prüferinnen neigten dazu, die erbrachte Leistung schlechter einzuschätzen als erfahrene Prüfer/innen.

Beide Prüfergruppen bewerteten die Soziale Kompetenz (Item 5), trotz identischer Standardisierung, bei den Borderline-Prüflingen schlechter als bei den exzellenten Prüflingen (siehe Tabelle 3). Der Unterschied war statistisch signifikant (4,80 vs. 4,13, $p<0,001$).

Akzeptanz für Standardisierten Prüflinge

Beide Prüfergruppen nahmen die Standardisierten Prüflinge als authentische Prüfungskandidaten wahr und sahen dieses neue Instrument als Möglichkeit einen OSCE weiter zu objektivieren. Beide Prüfergruppen empfanden die Einschätzung der Leistung bei guten Studierenden tendenziell leichter als bei den Borderline Studierenden, nahmen aber insgesamt keine Schwierigkeiten in der Leistungseinschätzung wahr.

Der regelmäßige Einsatz von Standardisierten Prüflingen zur Schulung von erfahrenen Prüfern/innen wurde von der Gruppe der unerfahrenen Prüfern mehr befürwortet als von den erfahrenen (2,9 vs. 2,0). Die detaillierten Ergebnisse sind in Abbildung 6 dargestellt.

Diskussion

Detaillierte Handlungsanweisungen für den Aufbau, die Umsetzung und Qualitätssicherungsmaßnahmen für einen OSCE und daraus resultierende gute teststatistische Ergebnisse, rechtfertigen den Einsatz dieses Prüfungsformats zur Überprüfung und Bewertung klinisch-praktischer Fertigkeiten an medizinischen Fakultäten [9], [15], [16], [17], [18]. Während OSCEs und OSPEs bislang vorwiegend in fakultätsinternen Prüfungen eingesetzt wurden, macht die aktuelle Diskussion zu ihrem Einsatz im Staatsexamen die Notwendigkeit einer Fakultätsübergreifenden Standardisierung deutlich [19]. Trotz etablierter Qualitätssicherungsmaßnahmen konnten potenzielle Einflüsse auf OSCE-Ergebnisse in verschiedenen Untersuchungen nachgewiesen werden. Dabei setzen solche Studien häufig ein hohes Maß an Personalaufwand, z.B. unabhängige Zweit-Bewerter, Video-Bewertungen, etc. voraus. Gleichzeitig lassen sich individuelle Einflussfaktoren von Prüflings- und Prüfer-Seite nicht eliminieren und nicht zufriedenstellend standardisieren. Unser Ziel war es durch die Anwendung des Konzepts der Standardisierung auf die studentische Leistung ein neues Instrument zur Qualitätssicherung in einem OSCE zu entwickelnd, das ermöglicht einzelne Einflussfaktoren auf die Bewertung der studentischen Leistung zu identifizieren. Gleichzeitig soll dieses neue Instrument zukünftig auch als Schulungs-Tool für OSCE-Prüfer eingesetzt werden können.

Im Rahmen der Überprüfung der Standardisierten Prüflinge konnte gezeigt werden, dass eine Standardisierung von Studierenden für ein zuvor definiertes Leistungsniveau gelingt.

Die Überprüfung der Standardisierung zeigte, dass bei beiden Prüflingsgruppen Abweichungen auftraten. Exzellente Prüflinge neigten eher dazu, zu wenige Antworten zu nennen und hatten Schwierigkeiten nicht einfach auswendig gelernte Listen wiederzugeben, wohingegen die Borderline-Prüflinge sowohl zu viele als auch zu wenig Antworten nannten. Die Abweichungen waren bei den Borderline-Prüflingen insgesamt deutlicher, was darauf hinweist, dass die Standardisierung für dieses Leistungsniveau schwieriger ist.

Tabelle 3: Prüfungs-Ergebnisse für Exzellente und Borderline Studierende aufgeschlüsselt nach Prüfererfahrung, Angaben als Mittelwerte (Min-Max)

	Erfahrene Prüfer (N=10)	Unerfahrene Prüfer (N=10)	p-Wert
Exzellente Studierende Gesamtpunktzahl	24,25 (20-25)	24,50 (23-25)	0,441
Item 1	4,80	4,85	0,687
Item 2	5,00	4,95	0,324
Item 3	4,95	5,00	0,324
Item 4	4,89	4,95	0,530
Item 5	4,85	4,75	0,442
Borderline Studierende Gesamtpunktzahl	15,15 (12-19)	13,50 (7-17)	**0,035**
Item 1	2,50 (1,539)	1,60 (0,598)	**0,020**
Item 2	2,15 (0,617)	2,00 (0,918)	0,559
Item 3	2,10 (0,718)	1,55 (0,686)	**0,018**
Item 4	2,50 (0,761)	4,35 (0,813)	**<0,001**
Item 5	4,25 (0,851)	4,00 (0,973)	0,393

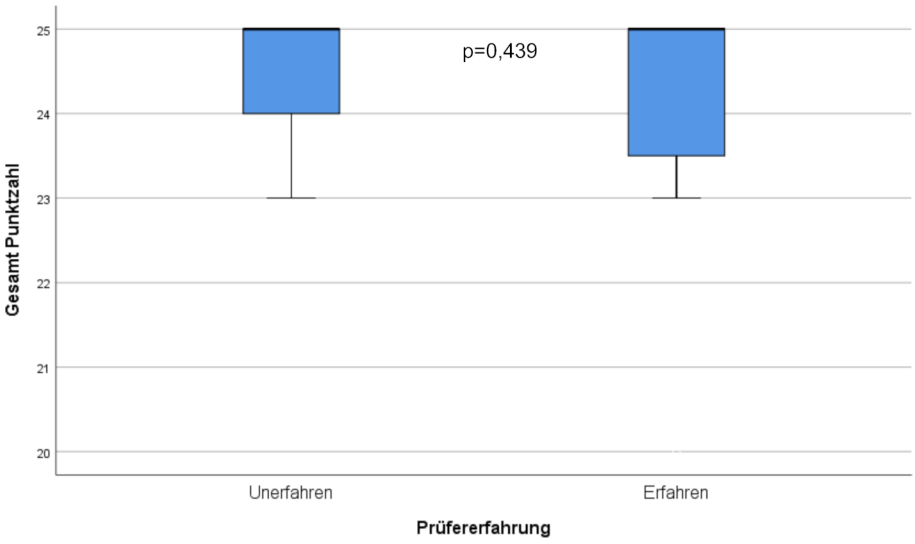

Abbildung 4: Gesamtpunktzahl für die Studierenden mit exzellenter Leistung nach Prüfererfahrung

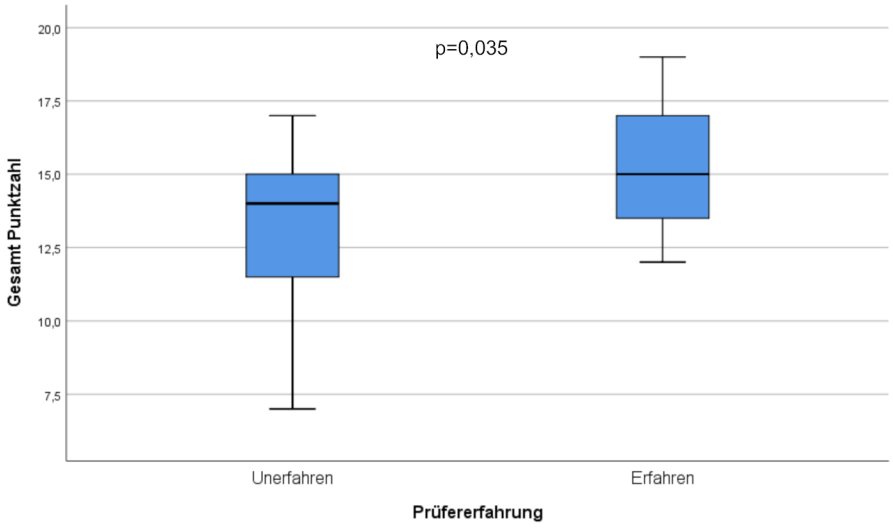

Abbildung 5: Gesamtpunktzahl für die Studierenden mit Borderline-Leistung nach Prüfererfahrung

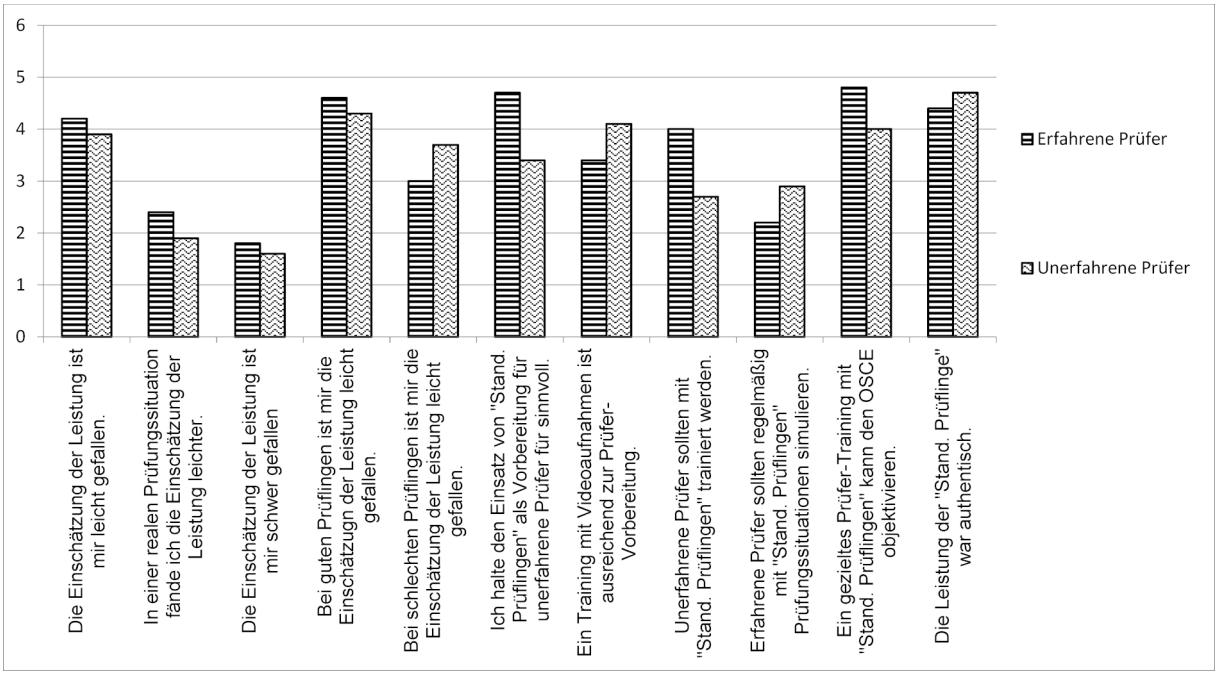

Abbildung 6: Ergebnisse der Evaluation der Standardisierten Studierenden durch erfahrene und unerfahrene OSCE-Prüfer/innen. Bewertung anhand einer 5 Punkte Likert-Skala (1= trifft gar nicht zu, 2=trifft wenig zu, 3=trifft teils zu, 4=trifft ziemlich zu, 5=trifft voll zu)

Borderline-Prüflinge wichen insbesondere bei Items von den Antworten ab, in denen die Darstellung eines diagnostischen oder therapeutischen Algorithmus gefordert war, siehe Tabelle 1. Dieser Umstand deutet darauf hin, dass möglicherweise eine zunehmende Komplexität der Aufgabe die Standardisierung erschwert. Analog verhielt es sich für die komplexeren Untersuchungsschritte aus der Checkliste *Abdominelle Untersuchung*. Hier wichen die Borderline-Prüflinge ebenfalls von den erwarteten Untersuchungsschritten ab (siehe Tabelle 1). Neben den rein inhaltlichen Abweichungen, neigten einzelne Studierende zur Überinterpretation der Rolle.

Sowohl die inhaltlichen Abweichungen als auch die unterschiedlichen Auslegungen der gespielten Rolle seitens der Standardisierten Prüflinge, weisen darauf hin, dass der Standardisierungs-Vorgang selbst und das Training der Rolle essentiell sind. In der hier gewählten Herangehensweise wurden die Studierenden anhand von modifizierten Checklisten, auf denen abhängig vom Leistungsniveau für jede mögliche Antwort definiert wurde, ob diese gegeben werden soll oder nicht, trainiert. Aus den Ergebnissen lässt sich ableiten, dass die Standardisierung noch genauer trainiert werden sollte. Zusätzlich erscheint es sinnvoll, analog zum Training von Simulationspatienten [26], zusätzlich eine vollständige Rolle zu definieren, in die dann das Leistungsniveau bzw. die zu überprüfende Eigenschaft eingebettet werden kann. Da die Prüfer/innen dazu neigten gerade bei den Borderline-Prüflingen Nachfragen zu stellen, müssen die Studierenden für solche Situationen ganz besonders geschult werden. Insbesondere muss dabei das Augenmerk auf komplexe Aufgebanstellungen und Untersuchungsgänge gelegt werden. Basierend auf den hier dargestellten Erfahrungen erscheint es sinnvoll, die Studierenden wiederholt ihr Rolle zur Überprüfung spielen zu lassen und dabei verschiedenen Optionen der Prüfer-Intervention zu simulieren um das Rollen-konforme Verhalten der Standardisierten Prüflinge zu üben. Eine Überprüfung der Standardisierung in einem Realitäts-nahen OSCE ist dabei eine weitere Option zur Überprüfung. Videoaufnahmen mit anschließender Analyse durch die Trainer und Standardisierten Prüflinge stellen eine weitere Trainings-Möglichkeit dar.

Ein offensichtlicher Nachteil der hier vorgelegten Studie ist die geringe Fallzahl. Es handelt sich um ein Pilotprojekt, das einer Machbarkeitsstudie entspricht. Zukünftig sollte die Standardisierung von Prüflingen mit mehr Studierenden erfolgen und in einer größeren Anzahl von Prüfungssituationen als der hier gewählten Zahl überprüft werden.

Im zweiten Studienabschnitt wurden die Videoaufnahmen der OSCE-Station *Management eines Patienten mit Sigmadivertikulitis* auf beiden Standardisierungsniveaus verwendet. Untersucht wurde, in wie weit sich die Prüfererfahrung auf die Bewertung der erbrachten Leistung auswirkt. Es wurde diese Station verwendet, da hier die Standardisierung am besten war.

Die Ergebnisse dieses Studienteils zeigen, dass die Leistung der Borderline-Prüflinge unterschiedlich durch die beiden Prüfergruppen eingeschätzt wurde. Unerfahrene Prüfer/innen bewerteten die Leistung signifikant schlechter und nutzten dabei auch eine größere Punkterange aus. Prinzipiell sind hierfür mehrere Erklärungen denkbar. Erfahrene Prüfer und Prüferinnen erkennen die erbrachte Leistung als solche und ordnen sie richtig ein. Auf der anderen Seite könnte dieses Ergebnis auch darauf hinweisen, dass erfahrenen Prüfer/innen für erkennbare Leistungsniveaus nicht die volle Bewertungsbreite nutzen und nur, wie von Iramaneerat beschrieben, einen einge-

schränkten Punktebereich verwenden [19]. Gleichzeitig könnte dieses Ergebnis auch darauf hindeuten, dass unerfahrene Prüfer/innen, unter Umständen unsicherer in der Einordnung schlechterer Leistungen sind und diese möglichicherweise übertrieben schlecht bewerten. Yeates und Kollegen zeigten in einer Arbeit wiederum, dass verschiedene Prüfer den Fokus in der Bewertung einer Leistung unterschiedlich setzen [27]. Die hier dargestellten Ergebnisse können daher auch ein Hinweis dafür sein, dass mit zunehmender klinischer oder Prüfungserfahrung, Schwerpunkte für die Punktevergabe unterbewusst anders gewählt werden. Nicht völlig ausschließen lässt sich, dass alle Prüfer und Prüferinnen hier einem *leniency error* unterliegen, der durch eine generelle Tendenz gekennzeichnet ist, Leistungen schlechter oder im anderen Extrem besser zu bewerten als sie eigentlich sind [13]. Gleichzeitig könnte auch, der von Yeates und Kollegen beschriebenen Effekt, eingetreten sein, dass eine Borderline-Leistung besonders schlecht bewertet wird, wenn sie direkt nach einer sehr guten Leistung beurteilt werden muss [7]. Wobei in dem hier gewählten Design die erste und die letzte Leistung in der Video-Sequenz jeweils einer Borderline-Leistung entsprach und damit nur einmal die von Yeates und Kollegen beschriebene Konstellation bestanden hätte.

Durch die schlechtere Bewertung der Sozialen Kompetenz bei Borderline-Prüflingen (4,80 vs. 4,13, $p<0,001$) trotz identischer Standardisierung und gleicher Performance in der Überprüfung der Standardisierung, lässt sich ein Halo-Effekt für beide Prüfergruppen vermuten. Die Ergebnisse dieser Studie suggerieren, dass im Sinne eines Halo-Effekts, wie von Iramaneerat und Kollegen beschrieben, die schlechtere inhaltliche Leistung zu einer Fehlwahrnehmung des Kommunikationsverhaltens führt [21]. Davon waren erfahrene und unerfahrene Prüfer/innen in gleichem Maße betroffen, was darauf hindeutet, dass auch eine umfangreiche Erfahrung als OSCE-Prüfer/in diesen Effekt nicht negieren kann.

Die detektierten Unterschiede in der Bewertung von Borderline-Prüflingen in Abhängigkeit von der Prüfererfahrung suggerieren, dass diese Effekte potentiell für das Bestehen oder Nicht-Bestehen einer OSCE-Station ausschlaggebend sein können. Letzteres verdeutlicht, dass eine gezielte Prüfer-Vorbereitung essentiell ist, gerade wenn OSCEs zukünftig im medizinischen Staatsexamen eingesetzt werden.

Ein weiterer Aspekt der generell bedacht und in Folgestudien weiter untersucht werden sollte, ist die Frage, ob es einen Unterschied im Bewertungsverhalten von erfahrenen Prüfern/innen gibt, abhängig davon, ob sie Erfahrung als OSCE-Prüfer/in besitzen oder nur eine längere klinische Erfahrung bzw. ob beides gegeben ist. Die erfahrenen Prüfer/innen in der aktuellen Studie hatten alle mehr als 5 Jahre klinische Erfahrung, die Erfahrung als OSCE-Prüfer variierte allerdings zwischen 2 und mehr als 5 OSCE-Prüfungseinsätzen. Da es sich bei der hier vorliegenden Studie um ein Pilotprojekt mit kleiner Fallzahl handelt, wurde dieser Aspekt nicht weiterverfolgt.

In der hier vorgestellten Studie stellt die Nutzung von Videos zur Durchführung einer solchen Untersuchung an sich dabei keine Neuerung dar. Vielmehr bieten Standardisierte Prüflinge aber zukünftig die Möglichkeit losgelöst von Videoanalysen analoge Untersuchung in einem OSCE mit standardisierten Prüflingen durchzuführen. Dabei ist denkbar, Standardisierte Prüflinge als „Qualitätsstandard“ in einem OSCE mitlaufen zu lassen. Die Art des Trainings für die Standardisierung muss zwingend noch weiter ausgearbeitet werden um Abweichungen zu minimieren. Zu überprüfen bleibt auch, ob eine Standardisierung eines Studierenden für mehrere Checklisten möglich ist.

Schlussfolgerung

Durch die Standardisierung von simulierten Prüflingen für definierte Leistungsniveaus, ergibt sich zukünftig die Möglichkeit Einflüsse auf das Bewertungsverhalten von Prüfern in OSCEs direkt zu analysieren. Im Rahmen von High Stakes Prüfungen, gerade auch im Hinblick auf den zukünftigen Einsatz von OSCEs im medizinischen Staatsexamen, stellen Standardisierte Prüflinge, neben der Qualitätssicherung, ein mögliches Instrument zur Schulung von OSCE-Prüfern dar [19].

Interessenkonflikt

Die Autor*innen erklären, dass sie keinen Interessenkonflikt im Zusammenhang mit diesem Artikel haben.

Literatur

1. Nikendei C, Kruppa E, Jünger J. Einsatz innovativer Lern- und Prüfungsmethoden an den Medizinische Fakultäten der Bundesrepublik Deutschland- eine aktuelle Bestandsaufnahme. Dtsch Med Wochenschr. 2009;134:731-732.
2. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J. 1975;22(1):447-451. DOI: 10.1136/bmj.1.5955.447
3. Schleicher I, Leitner K, Jünger J, Möltner A, Rüssler M, Bender B, Sterz J, Stibane T, König S, Frankenhauser S, Kreuder JG. Does quantity ensure quality? Standardized OSCE-stations for outcome-oriented evaluation of practical skills at medical faculties. Ann Anat. 2017;212:55-60. DOI: 10.1016/j.aanat.2017.03.006
4. Byrne A, Soskova T, Dawkins J, coombes L. A pilot study of marking accuracy and mental workload as measure of OSCE examiner performance. BMC Med Educ. 2016;16:191. DOI: 10.1186/s12909-016-0708-z
5. Wood TJ, Chan J, Humphrey-Murto S, Pugh D, Touchie C. The influence of first impressions on subsequent ratings within an OSCE station. Adv Health Sci Educ Theory Pract. 2017;22(4):969-983. DOI: 10.1007/s10459-016-9736-z
6. Fuller R, Homer M, Pell G, Hallam J. Managing extremes of assessor judgement within the OSCE. Med Teach. 2017;37(1):58-66. DOI: 10.1080/0142159X.2016.1230189

7. Yeates P, Cardell J, Byrne G, Eva KW. Relatively speaking: contrast effects influence assessors' scores and narrative feedback. Med Educ. 2015;49(9):909-919. DOI: 10.1111/medu.12777
8. Bartman I, Smee S, Roy M. A method of identifying extreme OSCE examiners. Clin Teach. 2013;10(1):27-31. DOI: 10.1111/j.1743-498X.2012.00607.x
9. Pell G, Fuller R, Homer M, Robert T. How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. Med Teach. 2010;32(10):802-811. DOI: 10.3109/0142159X.2010.507716
10. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81 Part I: A historical and theoretical perspective. Med Teach. 2013;35(9):e1437-1446. DOI: 10.3109/0142159X.2013.818634
11. Chesser A, Cameron H, Evans P, Gleland J, Boursicot K, Mires G. Sources of variation in performance on a shared OSCE station across four UK medical schools. Med Educ. 2009;43(6):526-532. DOI: 10.1111/j.1365-2923.2009.03370.x
12. Humphrey-Murto S, Touchi C, Wood TJ, Smee S. Does the gender of the standardised patient influence candidate performance in an objective structured clinical examination? Med Educ. 2009;43(6):521-525. DOI: 10.1111/j.1365-2923.2009.03336.x
13. Harasym PH, Woloschuk W, Cunning L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. Adv Health Sci Educ Theory Pract. 2008;13(5):617-632. DOI: 10.1007/s10459-007-9068-0
14. Turner JL, Dankosko ME. Objective structured clinical exams: A critical review. Fam Med. 2008;40(8):574-578.
15. Schultz JH, Nikendei C, Weyrich P, Möltner A, Fischer M R, Jünger J. Qualitätssicherung von Prüfungen am Beispiel des OSCE-Prüfungsformats: Erfahrungen der Medizinischen Fakultät der Universität Heidelberg. Z Evid Fortbild Qual Gesundhwes. 2008;102(10):668-672. DOI: 10.1016/j.zefq.2008.11.024
16. Barman A. Critiques on the objective structured clinical examination. Ann Acad Med Singapore. 2005;34(8):478-482.
17. Sloan DA, Donelly MB, Schwartz RW, Strodel WE. The Objective Structured Clinical Examination. The new gold standard for evaluating postgraduate clinical performance. Ann Sur.g 1995;222(6):735-742. DOI: 10.1097/00000658-199512000-00007
18. Mash B. Assessing clinical skill - standard setting in the objective structured clinical exam (OSCE). South Afr Fam Pract. 2007;49(3):5-7. DOI: 10.1080/20786204.2007.10873520
19. Jünger J. Kompetenzorientiert prüfen im Staatsexamen Medizin. Bundesgesundheitsbl. 2018;61:171-177. DOI: 10.1007/s00103-017-2668-9
20. Yeates P, O'Neill P, Mann K, Eva KW. 'You're certainly relatively competent': Assessor bias dur to recent experiences. Med Educ. 2013;47:910-922. DOI: 10.1111/medu.12254
21. Iramaneerat C, Yudkowsky R. Rater errors in a clinical skills assessmant of medical students. Eval Health Prof. 2007;30(3):266-283. DOI: 10.1177/0163278707304040
22. Schleicher I, Leitner K, Juenger H, Moeltner A, Ruesseler M, Bender B, Sterz J, Schuettler KF, Koenig S, Kreuder JG. Examiner effect on the objective structured cliniclal exam - a study at five medical schools. BMC Med Educ. 2017;17:71. DOI: 10.1186/s12909-017-0908-1
23. Nikendei C, Kraus B, Lauber H, Schrauth M, Weyrich P, Zipfel S, Jünger J. An innovative model for teaching complex clinical procedures: Integration of standardised patients into ward round training for final year students. Med Teach. 2007;29(2-3):246-252. DOI: 10.1080/01421590701299264
24. Rethans JJ, Grosfeld FJ, Aper L, Reniers J, Westen JH, van Wijngaarden JJ, van Weel-Baumgarten EM. Six formats in simulated and standardized patients use, based on experiences of 13 undergraduate medical curricula in Belgium and the Netherlands. Med Teach. 2012;34(9):710-716. DOI: 10.3109/0142159X.2012.708466
25. Barrows HS. An Overview of the uses of standardized patients for teaching and evaluating clinical skills. Acad Med. 1993;68(6):443-451. DOI: 10.1097/00001888-199306000-00002
26. Schulz JH, Schönemann J, Lauber H, Nikendei C, Herzog W, Jünger J. Einsatz von Simulationspatienten im Kommunikations- und Interaktionstraining für Medizinerinnen und Mediziner (Medi-KIT): Bedarfsanalyse - Training - Perspektiven. Gruppendyn Organisationsberat. 2007;38(1):7-23. DOI: 10.1007/s11612-007-0002-y
27. Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently - Mechanisms that contribute to assessor differences in directly-observed performance assessments. Adv Helath Sci Educ Theory Paract. 2013;18(3):325-341. DOI: 10.1007/s10459-012-9372-1

Korrespondenzadresse:
Petra Zimmermann
Ludwig-Maximilians-Universität München, Klinikum der Universität, Klinik für Allgemein-, Viszeral- und Transplantationschirurgie, Marchionini Str. 15, 81377 München, Deutschland, Tel.: +49 (0)89/4400-711239
petra.zimmermann@med.uni-muenchen.de

Bitte zitieren als
Zimmermann P, Kadmon M. Standardized examinees: development of a new tool to evaluate factors influencing OSCE scores and to train examiners. GMS J Med Educ. 2020;37(4):Doc40. DOI: 10.3205/zma001333, URN: urn:nbn:de:0183-zma0013336

Artikel online frei zugänglich unter
https://www.egms.de/en/journals/zma/2020-37/zma001333.shtml

Eingereicht: *15.10.2019*
Überarbeitet: *23.02.2020*
Angenommen: *27.04.2020*
Veröffentlicht: *15.06.2020*