Formative assessment of practical skills with peer-assessors: quality features of an OSCE in general medicine at the Heidelberg Medical Faculty

Abstract

Background: Objective Structured Clinical Examinations (OSCEs) have become an established examination format at German medical faculties. Medical experts routinely use a summative assessment to evaluate practical and communicative skills, while the use of the OSCE format by student examiners, as a formative examination, remains rather limited.

Objective: The formative OSCE program of the Department of General Practice and Implementation Research at the Heidelberg Medical Faculty, which is conducted and evaluated by peer tutors, is examined with regard to its quality criteria and compared with summative OSCEs from other departments.

Methods: Difficulties and discriminatory power of individual testing stations were determined for the summative, as well as the formative OSCE, and compared with each other. To assess the reliability of the measurements, an analysis of the data was carried out using the Generalizability theory. In addition, a comparison is made between the assessments of student examiners and second assessments by medical experts.

Results: The stations of the formative OSCE show similar difficulties as those of the summative comparison OSCEs (P_{form}=0.882; P_{sum}=0.845 – 0.902). With respect to measurement reliability, there are no differences between the OSCE in General Medicine and the other subjects. The assessments of student examiners and medical experts correlate highly (r=0.888).

Conclusion: The formative OSCE in General Medicine is comparable to the summative comparison formats in terms of its quality criteria. The use of student examiners can be a reliable alternative to medical experts in formative OSCEs.

Keywords: formative, OSCE, student examiners, generalizability theory

Andreas Möltner[1]
Mirijam Lehmann[1]
Cornelia Wachter[2]
Sonia Kurczyk[2]
Simon Schwill[2]
Svetla Loukanova[2]

1 University Heidelberg, Baden-Württemberg Center of Excellence for Assessment in Medicine, Heidelberg, Germany

2 University Heidelberg, Medical Faculty, Department of General Practice and Implementation Research, Heidelberg, Germany

1. Introduction

Practical clinical skills and anamnesis are already being taught at various medical faculties in the preclinical study semesters and tested with the help of an Objective Structured Clinical Examination (OSCE). It has been shown that an early learning of practical skills leads to better results in the clinical examination sections and clinical skills [1].

Traditionally, the teaching content is taught by faculty physicians, but increasingly also by student tutors of higher semesters. An advantage of peer tutors (Peer Assisted Learning, PAL) is the higher acceptance by students [2], lower costs [3], [4] and the possibility of smaller learning groups [5]. In addition, the students benefit from a reduction of stress and anxiety factors [6] and the student tutors [2], [7] benefit from their own in-depth study of the learning content. When comparing the student tutors with faculty members, the PAL students achieve the same results in final exams [8], [9], [10], [11] and the same or even higher quality of feedback [10]. Prerequisites for this are precisely defined student tutor training courses and checklists [12], [13].

Since 2013, practical skills and anamnesis techniques have been taught at the Heidelberg Medical Faculty in the pre-clinical part of the AaLPLUS courses (AaL: “Living Anatomy Plus”) of the Department of General Practice and Implementation Research with the help of peer tutors and subsequently examined in a formative OSCE, also conducted by student tutors [14]. A detailed description of the program and the evaluation of the OSCE by students and peer tutors can be found in [15].

Black and Wiliam [16] see five essential aspects of formative examinations. These are adapted to the context of University education:

1. Clarification and exchange of learning goals and success criteria
2. Initiating effective discussions and other learning tasks that demonstrate students' understanding of the learning content
3. Feedback that is useful for the students
4. Encouraging students to act as a mutual learning resource
5. Encouraging students to see themselves as initiators of their own learning activities

These objectives involve a whole process of teaching in which more or less continuously formative examinations are integrated. This is often logistically difficult to achieve fully in formative practical examinations in the form of OSCEs in medical education, so that the formative OSCE considered here should rather be seen as an instrument [17], which comes at the end of the pre-clinical part of the study. In order to achieve the goals announced by Black and Wiliam, other forms of formative examination procedures should be suitable [18]. Despite this limited function of the formative OSCE, it can be expected to have a positive effect on the learning behaviour of the examined students [19], [20].

In a review article by Khan et al. from 2017, 13 publications on the topic of "Students as examiners in OSCEs" are presented in more detail [21]. Some of the papers listed there examine the assessments of students and experts with regard to basic characteristics such as differences in the scores awarded and the correlation of the assessments of students and experts as examiners. A more detailed quantitative analysis, which also includes a differentiation of station- and examiner effects and their consequences for measurement reliability, is only provided in the works of Moineau et al. [10] and Basehore et al. [22]. In both studies, double evaluations at the stations by students and experts are investigated (in [22] the experts evaluated using videos of the examinations). However, it was not investigated whether student examiners differ from experts with regard to the extent of exam effects.

Besides the comparison of student examiners and experts in the same formative examination, the quality of the formative examination in relation to the summative examinations established at the faculty is also of interest. Formative examinations differ in their objectives and structure (e.g. higher importance of feedback) and relevance of summative examinations to the examined students. The latter in particular, can have an effect on the reliability and accuracy of measurements, e.g. if the performance of the candidates is less differentiated due to reduced motivation.

Aim of the study

The aim of the study was to demonstrate

1. that students in the context of formative examinations of practical skills can replace experts as examiners without compromising the quality of the examination and
2. that the quality of such formative examinations reaches the same standards as established summative examinations.

To this end, the formative OSCE in General Medicine at the Heidelberg Medical Faculty, which was held in 2018 and involved tutors as examiners, was examined with regard to its quality criteria (characteristics of the stations, measurement reliability of the exam, extent of examiner effects). A comparison was made with summative OSCEs, and a matching between the assessments of student examiners and those of experts ("supervisors") was considered.

Other aspects of the formative OSCE in General Medicine with student examiners, such as acceptance by both examiners and examined students, assessment of the quality of feedback and subjective benefit to both students and examiners of the skills assessed in the OSCE are described in detail in [15]. The present study focuses exclusively on the quality characteristics of the OSCE that can be measured by statistical parameters of the examination results.

Standard analyses of tests usually include basic parameters such as difficulty, selectivity and reliability (see 3.1.1). Based on the Generalizability theory, the facets (influencing factors) "students" (differences in the ability of students), "station" (difference in the difficulty of stations), "examiner" (difference in the "strictness" of examiners) and the interaction "station x examiner" (different strictness of examiners at different stations) and their effects on generalizability and absolute measurement accuracy (see 3.1.2) were examined.

To compare the characteristic values of the OSCE General Medicine with established summative OSCEs of the Heidelberg Medical Faculty, the OSCEs of the subjects Surgery and Internal Medicine of the winter semester (winter term) 2017/2018, the summer semester (summer term) 2018 and the winter term 2018/2019 were used. Finally, a comparison of double assessments by student examiners and experts within the formative OSCE General Medicine was conducted (3.2).

2. Methods

2.1. Implementation of the OSCE

The formative OSCE General Medicine in May 2018 was attended by 300 students of the fourth semester. The OSCE took place over two days and comprised four testing stations. One of the four stations ("venous blood sampling") was completed by all students. Various clinical examinations had to be performed at two stations. These stations were not identical for the participating students, but alternated between the different parcours. A total of 11 different tasks were used (general examination of the abdomen, examination of spleen/kidney/appendicitis signs, blood pressure measurement, examination of the

Table 1: Number of assessments in the OSCE General Medicine 2018 by station (PE: Physical Examination).

Station	Ratings	Supervisions
Anamnesis Abdomen	66	7
Anamnesis Head	94	8
Anamnesis Back	140	22
PE Abdomen I: General examination of the abdomen	73	9
PE Abdomen II: Spleen, Kidney and Signs of Appendicitis	47	0
PE Blood Pressure Measurement	46	0
PE Heart	47	12
PE Liver	24	8
PE Lymph Notes	60	8
PE Neurology	60	24
PE Pulse Measurement	71	8
PE Thyroid Gland	50	8
PE Thorax/Lungs	50	8
PE Spinal Cord	72	6
Venous Blood Sampling	300	7
	1200	135

heart, liver, lymph node status, pulse status, thyroid gland, thorax, spine and a neurological examination). Furthermore, a complete anamnesis had to be taken. Here too, the contents changed (back, abdomen and head). Trained acting patients were used for the clinical examinations and the anamnesis. The contents of the stations and the essential criteria for evaluation were known to the participating students from the previous tutorials and given materials.

Each participant went through a total of four stations of eight minutes duration (5 minutes per task and 3 minutes feedback). The assessment of performance was carried out using checklists by students with basic didactic training who were at least in their sixth semester. A total of 25 points could be achieved at each of the stations. An exception to this were the three stations where an anamnesis had to be taken. In these, 30 points were to be achieved.

32 students were used as examiners, 26 of whom examined at several (up to five) stations during the course of the OSCE (see table 1). The assessments were recorded using tablet computers (Programm tOSCE des UCAN-Prüfungsverbunds) [23].

Five supervisors were appointed to monitor the quality of implementation and evaluation, who carried out random second evaluation (135 evaluations in total). The trained examiners were (medical) staff members of the Department of General Practice and Implementation Research and, for the assessment of communicative skills at the three anamnesis stations, lecturers of the Department of Medical Psychology.

2.2. Comparison with summative OSCEs

Six OSCEs of the subjects Surgery and Internal Medicine of the winter semesters 2017/2018 and 2018/2019 and of the summer semester 2018 of the Heidelberg Medical Faculty were used to compare the quality criteria of the OSCE General Medicine. The inclusion of several comparative OSCEs from two different subjects and semesters ensures that an estimate of the variability of their characteristic values (e.g. proportion of examiner influences) can be made for the comparative OSCEs.

The OSCEs in Internal Medicine comprised 10 stations, those in surgery 13 stations. A maximum of 25 points could be achieved at all stations of these OSCEs (see table 2). These OSCEs were performed on two to three days in two parallel courses (viz. “parcours”). The stations were partly changed in the different parcours. The two subjects Internal Medicine and Surgery were chosen because:

1. different examiners were used at the same testing stations and
2. the examiners were generally employed at different stations.

This allows an estimation of the examiner, stations and the interaction effect station x station during the evaluation.

2.3. Statistical analysis

Difficulty P and corrected selectivities r_{it} (correlations of the number of points achieved at one station with the points achieved at all other stations) as well as the mean inter-correlations with all other stations r_{ij} (mean inter-item correlation) were determined for the stations of all mentioned OSCEs. The product-moment correlation (according to Pearson, two-tailed P value) was used throughout as a correlation measure.

In order to achieve equivalence of the stations, the point values obtained at the anamnesis stations, where 30 points were to be achieved, were rescaled to the range of 0-25 points for all analyses presented.

To assess the reliability of the measurements, the data were analysed using the Generalisability theory [24]. The facets considered were “students”, “stations”, “examiners” and the interaction “station x examiner”. From the

Table 2: Number of participants, stations and examiners in the OSCE General Medicine and the OSCEs Surgery and Internal Medicine WS2017/2018 to 2018/2019.

OSCE	Participants	Stations	Examiners
General Medicine 2018	300	15	32
Internal Medicine WS18	160	28	24
Surgery WS18	173	52	58
Internal Medicine SS18	193	28	26
Surgery SS18	182	52	56
Internal Medicine WS17	179	27	30
Surgery WS17	145	47	63

variance components found by applying the Generalizability theory, the "generalizability" $E\rho^2$ (as an analogy to internal consistency/Cronbachs α) and the "dependability" Φ can be determined as a measure of absolute measurement accuracy:
If n denotes the number of stations, then

$$E\rho^2 = \frac{\sigma^2_{Stud}}{\sigma^2_{Stud} + \sigma^2_{Resid}/n}$$

$$\theta = \frac{\sigma^2_{Stud}}{\sigma^2_{Stud} + (\sigma^2_{Station} + \sigma^2_{Prüfer} + \sigma^2_{Station*Prüfer} + \sigma^2_{Resid})/n}$$

In order to analyse the matching between the assessments of the student examiners and the supervisors, the scores awarded for each station were compared (Wilcoxon signed-rank test) and the correlations determined. Furthermore, an analysis of variance of the total data set (examiners and supervisors) with the fixed factor "student examiner/supervisor" and the facets "students", "stations", "student examiners", "supervisor" and the interaction "station x examiner" was carried out.
Note: When analysing with the Generalizability theory, a distinction must be made between so-called fixed and random factors. If the facet "student" is considered a random factor, the intention is to generalise to equivalent groups of students (i.e. in the same semester, same demographic composition, equivalent teaching etc.). The group of students considered in the examination being analysed should therefore be regarded as a sample from a population. The same applies to the facet "station": As a random factor, the focus is on generalizability to equivalently constructed stations, while the facet "examiners" involves examiners from a potential group of examiners. When modelling the station or examiner as a fixed factor, however, the focus is on the stations or examiners actually used in the exam: Are individual stations particularly easy or difficult, are examiners too strict or too lenient? Since the present study focuses on generalizability, only the results for the analyses with "student", "station" and "examiner" are presented as random factors.
The statistical analyses were performed with R Version 3.5.1. For the mixed model analyses for evaluation with the model of generalizability theory the packages "lme4" and "lmerTest" were used.

3. Results

3.1. Characteristic values of the test

3.1.1. Difficulties and selectivity of testing stations

The basic parameters (mean score achieved x, difficulty P and corrected selectivity rit) of the scores obtained at the stations are listed in table 3. Figure 1 contains a graphical representation of the distributions as a box plot. The difficulties at the individual stations range from P=0.794 at the "Anamnesis Abdomen" station to P=0.959 at the "Blood Pressure Measurement" station. An average of 87.632 out of a maximum of 100 points was achieved. Please note that in contrast to dichotomous items, where only 0 or 1 point can be achieved, with finer granular evaluations (here 0-25 points) selectivities can possibly be interpreted even if the difficulties are numerically high. Eleven of the 15 stations have part-hole corrected selectivities of more than 0.300, two stations are just below this limit with selectivities of 0.276 and 0.296 ("Physical Examination Blood Pressure" and "Physical Examination Neurology"). Significantly lower are the stations "Physical Examination Liver" with r_{it}=0.112 and "Pulse status" with r_{it}=0.099.

Comparison with summative OSCEs

Figure 2 shows the distribution of the scores achieved at the stations of OSCE General Medicine compared to the summative OSCEs in Internal Medicine and Surgery in the last three semesters (see also table 4).
In comparison to the considered OSCEs of Internal Medicine and Surgery, the stations of the OSCE General Medicine were almost equally heavy (P=0.882 compared to P=0.876).
The corrected selectivities were on average lower than in the comparative OSCEs, only the OSCE Internal Medicine SS 2018 showed lower values (r_{it}=0.358 compared to 0.386, see table 4 and figure 3). In this comparison, however, it must be taken into account that in the OSCE General Medicine, the point total of the other stations used for the corrected selectivity is determined from only three stations, in contrast to Internal Medicine and Surgery with nine and twelve stations, respectively. This means that this sum is subject to more error variance in the OSCE General Medicine. A better possibility for com-

Table 3: Characteristic values of the stations of the formative OSCE General Medicine 2018.

Station	*N*	$\bar{x}$	*P*	r_{it}
Anamnesis Abdomen[1]	66	19.861	0.794	0.384
Anamnesis Head[1]	94	20.372	0.815	0.392
Anamnesis Back[1]	140	20.577	0.823	0.357
PE Abdomen I	73	21.863	0.875	0.541
PE Abdomen II	47	23.064	0.923	0.663
PE Blood Pressure Measurement	46	23.978	0.959	0.276
PE Heart	47	22.234	0.889	0.512
PE Liver	24	22.458	0.898	0.112
PE Lymph Nodes	60	22.767	0.911	0.451
PE Neurology	60	21.783	0.871	0.296
PE Puls Measurement	71	22.944	0.918	0.099
PE Thyroid gland	50	22.800	0.912	0.407
PE Thorax	50	21.180	0.847	0.485
PE Spinal column	72	22.361	0.894	0.419
Venous Blood Sampling	300	22.350	0.894	0.399

[1] Point values rescaled to the range 0-25

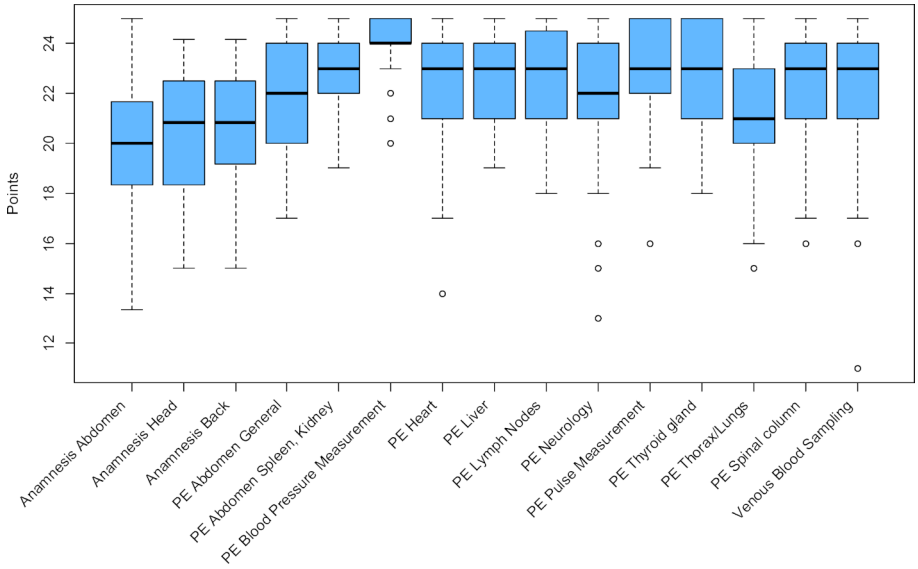

Figure 1: Distribution of the scores achieved at the stations of the formative OSCE General Practice. The station "Complete anamnesis", where 30 points were achieved in the original OSCE, has been rescaled to the range of 0-25 points.

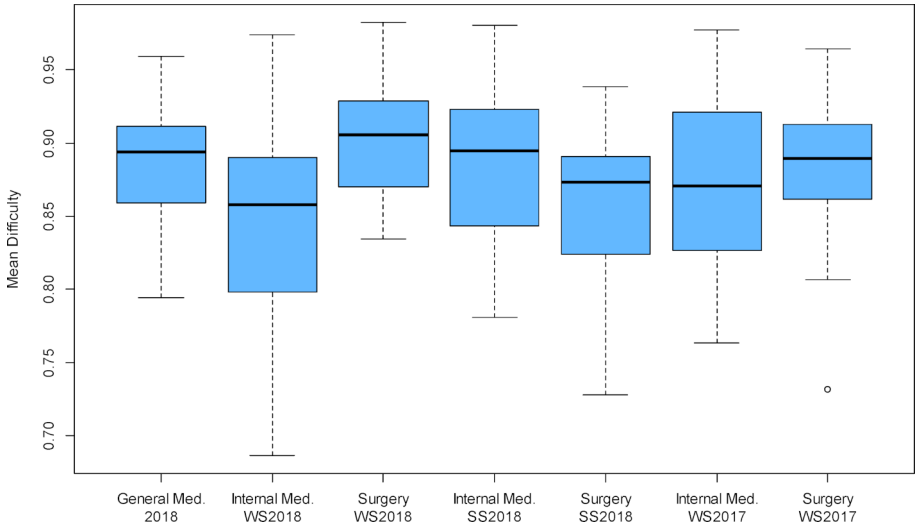

Figure 2: Distribution of the mean difficulties achieved P at the stations of the formative OSCE General Medicine 2018 and the summative OSCEs Internal Medicine and Surgery winter semester 2017/18 to winter semester 2018/2019.

Table 4: Average difficulties, discriminatory power and intercorrelations with other stations of the OSCE General Medicine and the OSCEs Internal Medicine and Surgery of WS 2017 – WS 2018.

OSCE	P	r_{it}	r_{ij}
General Medicine 2018	0.882	0.386	0.285
Internal Medicine WS18	0.845	0.535	0.359
Surgery WS18	0.902	0.417	0.233
Internal Medicine SS18	0.884	0.358	0.188
Surgery SS18	0.858	0.553	0.356
Internal Medicine WS17	0.871	0.487	0.292
Surgery WS17	0.883	0.461	0.266

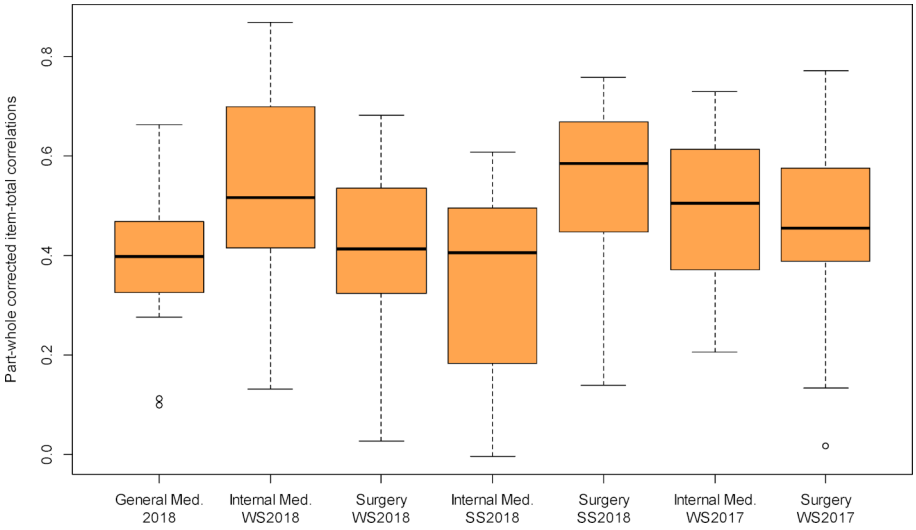

Figure 3: Distribution of corrected item-total correlations rit at the stations of the formative OSCE General Medicine 2018 and the summative OSCEs Internal Medicine and Surgery winter semester 2017/18 to winter semester 2018/2019.

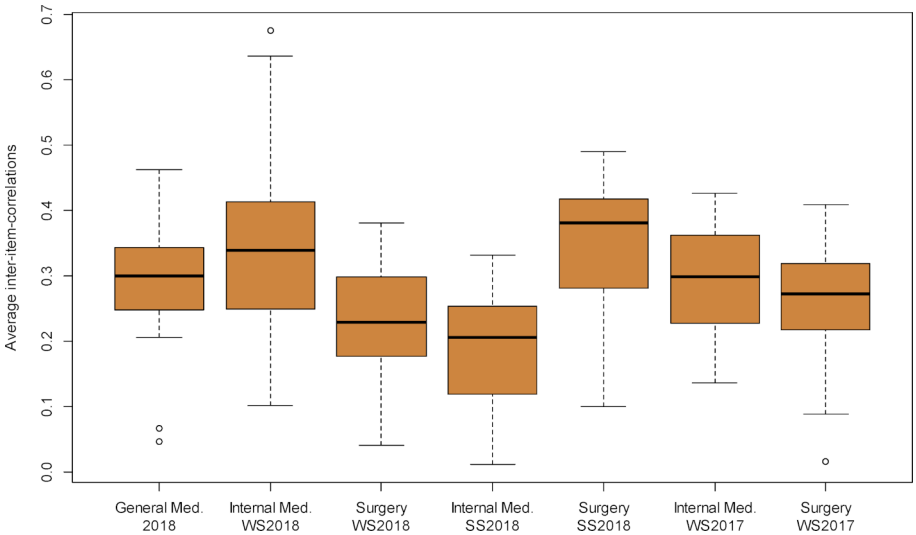

Figure 4: Distribution of averaged inter-item correlations r_{ij} (correlations of the number of points achieved at one station with the respective other stations) of the formative OSCE General Medicine 2018 and the summative OSCEs Internal Medicine and Surgery winter semester 2017/18 to winter semester 2018/2019.

parison is offered here by the average of all correlations of the point sum from one ward with all other stations r_{ij} ("mean inter-item correlation"). Here it can be seen that three of the comparison OSCEs each have lower and higher values (see table 4 and figure 4).

3.1.2. Measurement reliability

Methods of Generalizability theory were used to analyse measurement reliability. A model with the facets "student", "station", "examiner" and the interaction "station x examiner" was analysed. Table 5 shows the estimated variance components of the facets.

Table 5: Variance components for the facets of the student, station, examiner and station x examiner. The standard deviation indicates the size of the influence of the respective effect in points at a station.

Facets	Variance	Std.-Dev.	Variance [%]	*p*
Student	1.293	1.137	21.64	<0.001
Station	1.253	1.119	20.96	<0.001
Examiner	0.544	0.738	9.10	<0.001
Station x Examiner	0.060	0.246	1.01	0.337
Residuum	2.825	1.681	47.28	

Table 6: Estimated generalizability $E\rho^2$ and dependability Φ for the OSCE in General Medicine and the OSCEs in Internal Medicine and Surgery, assuming a parcour of 10 stations.

OSCE	$E\rho^2$	Φ
General Medicine 2018	0.821	0.734
Internal Medicine WS18	0.851	0.780
Surgery WS18	0.765	0.689
Internal Medicine SS18	0.751	0.668
Surgery SS18	0.861	0.807
Internal Medicine WS17	0.830	0.767
Surgery WS17	0.818	0.750

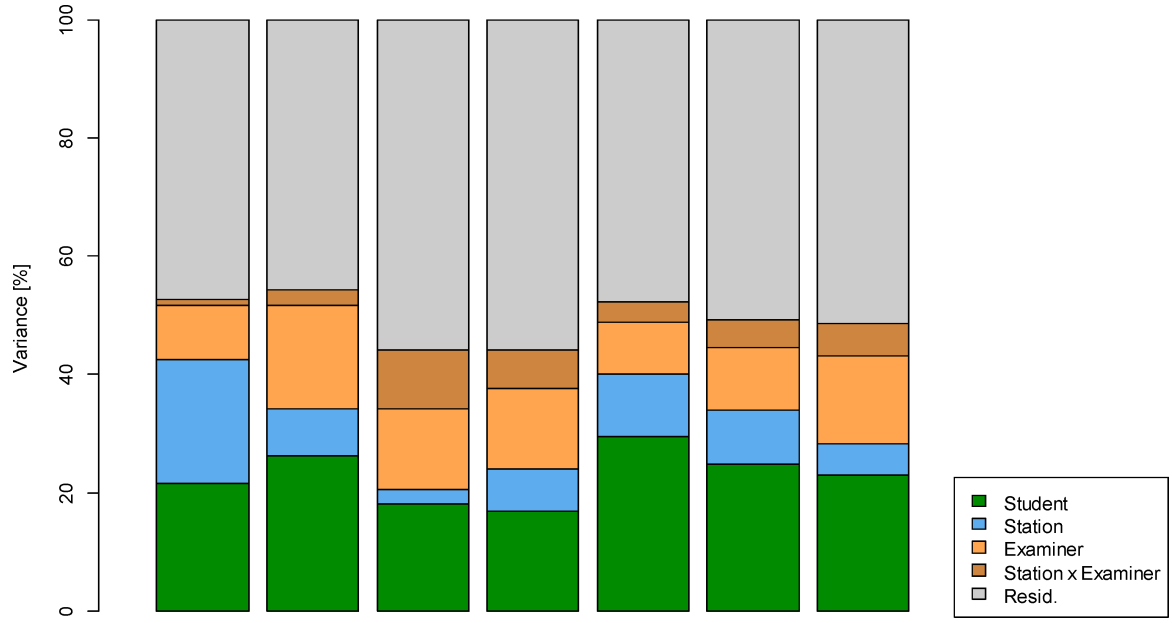

Figure 5: Percentage distribution of the variance of the OSCE General Medicine and the OSCEs Internal Medicine and Surgery from WS 2017 to WS 2018. The total variance is divided into the components "student", "station", "examiner", the interaction "station x examiner" and the residual variance.

Nearly 53% of the variance can be explained by the effects of the model, with 22% attributable to differences between students in terms of performance. The variability of the stations accounts for 21%, while the combined examiner influence was around 10%. The interaction effect station x examiner was not detectable or significantly different from 0.

The expected correlation of the point values achieved in the OSCE with an equivalent OSCE is $E\rho^2$=0.647. These values do not take into account the effects of station and examiner, since in an equivalent parcours, all students pass through the same stations with the same examiners, so their total achieved points are only changed by these facets by a value that is constant for all and is not taken into account in a correlation ($E\rho^2$ is thus a measure of the relative measurement accuracy). In contrast, the Dependability Φ, as a measure of the absolute measurement accuracy, takes these factors into account, and is Φ=0.525 for the test.

Comparison with summative OSCEs

Figure 5 shows graphically the percentage shares of the variance components for the OSCEs. A quality comparison of the OSCE General Medicine with those of Internal Medicine and Surgery with regard to the quality of the stations and the extent of the examiner's influences must take into account the different number of stations. As an example, table 6 lists the values obtained on a Parcour with ten stations. It can be seen that for $E\rho^2$ three of the

Table 7: Comparison of assessments by examiners and supervisors: mean scores of examiners (x), mean scores of supervisors (xS), significance value of test for difference (Wilcoxon signed-rank test, p) and correlation of assessments (r).

Station	*n*	$\bar{x}$	$\bar{x}_S$	*p*	*r*
Anamnesis Abdomen	7	20.595	19.643	0.034	0.989
Anamnesis Head	8	18.854	18.229	0.389	0.926
Anamnesis Back	22	18.750	18.902	0.483	0.729
PE Abdomen I	9	22.667	22.889	0.586	0.841
PE Abdomen II	12	21.083	21.500	0.120	0.946
PE Blood Pressure Measurement	8	21.500	21.375	0.773	0.917
PE Heart	8	21.500	22.000	0.265	0.783
PE Liver	24	22.250	22.292	0.903	0.904
PE Lymph Nodes	8	23.875	24.250	0.149	0.787
PE Neurology	8	21.500	22.125	0.269	0.851
PE Pulse Measurement	8	20.500	20.125	0.345	0.965
PE Thyroid gland	6	21.500	22.000	0.149	0.986
PE Thorax	7	21.857	21.857	1.000	0.917

Table 8: Variance components of the analysis of the assessments by student examiners and supervisors.

Facets	Variance	Std.-Dev.	p
Student	1.392	1.180	<0.001
Station	1.266	1.125	<0.001
Tutor (Examiner)	0.490	0.700	<0.001
Supervisor (Examiner)	0.311	0.557	0.117
Station x Examiner	0.106	0.325	0.079
Residuum	2.716	1.648	

six comparison OSCEs have both lower and higher values. The absolute accuracy is higher for four comparison OSCEs. As can be seen in figure 5, this is mainly due to the higher variability of the stations.

3.2. Supervision

In 135 assessments, an additional examination was carried out by a supervisor (medical staff of the Department of General Practice and Implementation Research and Medical Psychology), which serves as quality assurance of the OSCE (see table 1). Table 7 shows the mean values of the assessments by the examiners, as well as those of the supervisors for the wards with double assessments. In addition, the significance value of the test for difference of assessments (Wilcoxon signed-rank test) is given. Only one station ("Anamnesis Abdomen") shows a statistically significant difference.

Table 7 contains the correlations between examiners and supervisors at the stations, these ranged from 0.729 to 0.989. As examples, the scatter plots (bubble chart) of the assessments for the wards "Back Anamnesis" and "Physical Examination Neurology" are shown in figure 6.

An overall analysis based on the Generalizability theory of all data (student examiners and supervisors) with the examiner group as a fixed factor and with separate variance components for the two examiner groups is shown in table 8. The supervisors give 0.568 points less than the student examiners, but the effect is not significant (p=0.152). The examiner effects have a standard deviation of 0.700 points (see also table 5). For the five supervisors, no variance component other than zero can be demonstrated (p=0.117), which is equivalent to the fact that no difference can be demonstrated with regard to their strictness.

4. Discussion

The results show that the stations of the OSCE General Medicine 2018 essentially fulfill the same quality criteria as the stations that are tested in the OSCEs in the subjects of Surgery and Internal Medicine, which have been established for years. In two of the physical examination stations, a review is recommended due to low selectivity. The matching of the assessments of the student examiners with those of the supervisors can be described as good to very good at all stations. Systematic differences between the assessments of the student examiners and the supervisors cannot be demonstrated. Although there is a relative influence of the examiners, the examiner effects tend to be even lower than in the comparison OSCEs.

The generalizability standardized on ten stations is noticeably higher in the OSCE General Medicine with $E\rho^2$=0.82 compared to the two studies mentioned in Khan's review [21], in which an analysis was carried out using the Generalizability theory, in [10] and marginally higher in [22] ($E\rho^2$=0.51 for the checklist and $E\rho^2$=0.63 for the "global score" and $E\rho^2$=0.80 for the "total score").

Apart from the number of stations, the measurement reliability of the OSCE examination in General Medicine is fully in line with the summative comparative OSCEs in

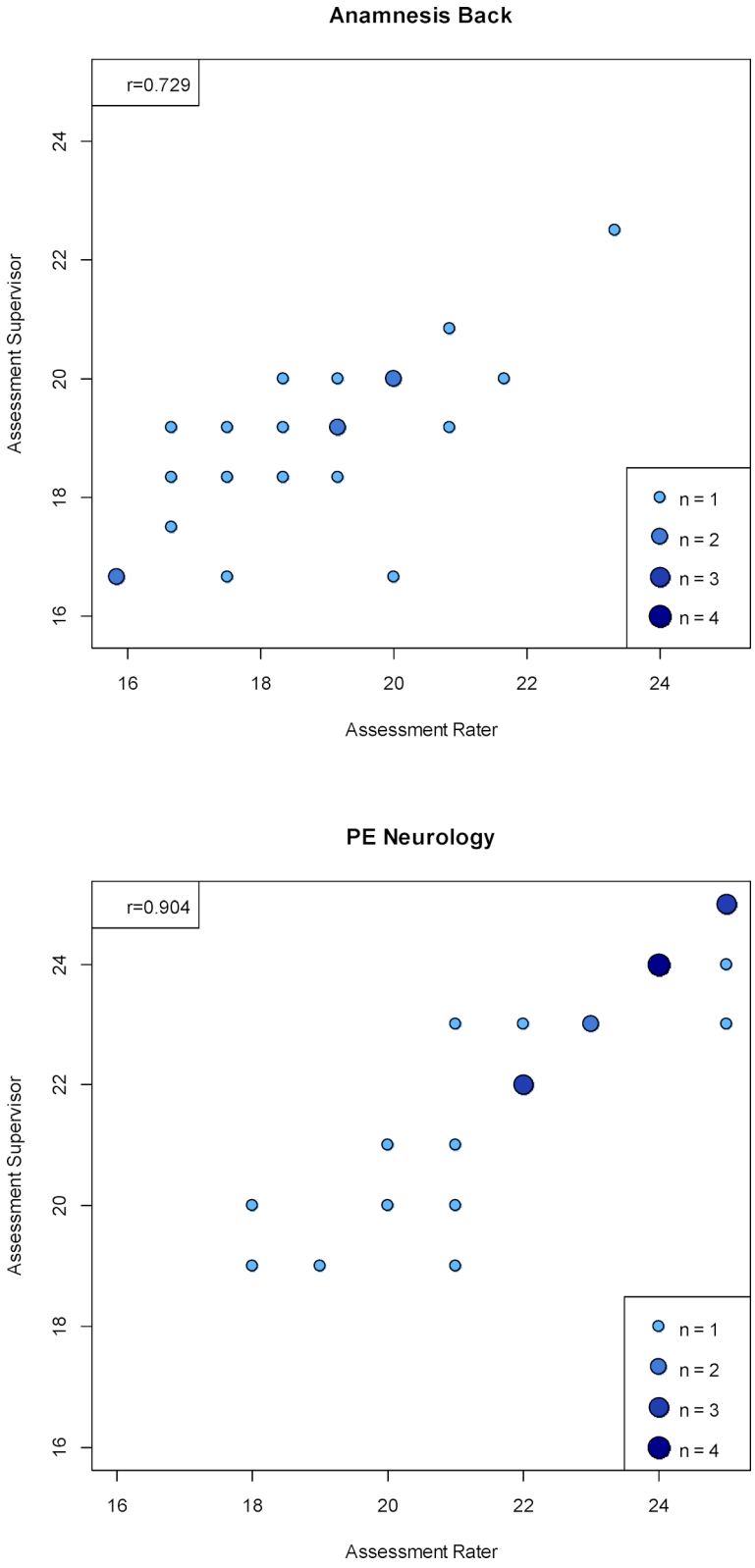

Figure 6: Scatter plots (bubble plots) of the assessments by examiners and supervisors at the "Back Anamnesis" and "Physical Examination Neurology" stations (the circle size represents the number of multiple data points with the same values).

the subjects of Surgery and Internal Medicine in the last three semesters.
This shows that with appropriate preparation:

1. students instead of experts can be used as examiners of practical skills and
2. the quality of a formative examination with student examiners is similar to that of established summative OSCEs with experts as examiners.

Since the implementation of practical format-based exams, which record the level of knowledge for students themselves as well as for teachers in a structured manner, often fails at the faculties due to the availability of examiners from the teaching staff, students in higher semesters offer a convenient alternative to substitute them.
The only weakness of the OSCE General Medicine is the small number of four stations that the examined students have to pass through. However, the fact that four stations does not provide a measurement reliability that meets the requirements of high-quality examinations is not surprising. This is in line with the literature, which demands significantly higher numbers of stations for OSCEs in order to obtain overall evaluations that can be classified as meaningful [25].
The analysis of other formative examinations in which students act as examiners is of course desirable, since it is not possible to generalize to other institutions, general conditions, or the like from the individual case presented here. Such investigations could show which conditions must be met for the use of student examiners in order to obtain statistically satisfactory and meaningful performance assessments.

Limitations

The random second assessment by the supervisors were not carried out systematically, so that the comparisons with the student assessors are partly based on very small data sets (see table 7). There is also a room for improvement in the systematic allocation of the two physical examination stations from the set of eleven available stations among the examined students.

5. Conclusion

Overall, the OSCE General Medicine shows that it is possible to assess a large number of students with student examiners and thus to conduct high quality formative practical examinations. The involvement of students in the process of creating formative performance assessments is thus a practical way for medical faculties to take advantage of the widely recognized benefits of feedback in university teaching with the help of structured performance recording.

Funding

The work was developed within the framework of the project MERLIN II (01PL17011C) funded by the Federal Ministry of Education and Research.

Competing interests

The authors declare that they have no competing interests.

References

1. Swierszcz J, Stalmach-Przygoda A, Kuzma M, Jablonski K, Cegielny T, Skrzypek A, Wieczorek-Surdacka E, Kruszelnicka O, Chmura K, Chyrchel B, Surdacki A, Nowakowski M. How does preclinical laboratory training impact physical examination skills during the first clinical year? A retrospective analysis of routinely collected objective structured clinical examination scores among the first two matriculating classes of a reformed curriculum in one Polish medical school. BMJ Open. 2017;7(8):e017748. DOI: 10.1136/bmjopen-2017-017748
2. Khalid H, Shahid S, Punjabi N, Sahdev N. An integrated 2-year clinical skills peer tutoring scheme in a UK-based medical school: perceptions of tutees and peer tutors. Adv Med Educ Pract. 2018;9:423-432. DOI: 10.2147/AMEP.S159502
3. Bosse HM, Nickel M, Huwendiek S, Schultz JH, Nikendei C. Cost-effectiveness of peer role play and standardized patients in undergraduate communication training. BMC Med Educ. 2015;15:138. DOI: 10.1186/s12909-015-0468-1
4. Lee CB, Madrazo L, Khan U, Thangarasa T, McConnell M, Khamisa K. A student-initiated objective structured clinical examination as a sustainable cost-effective learning experience. Med Educ Online. 2018;23(1):1440111. DOI: 10.1080/10872981.2018.1440111
5. Hudson JN, Tonkin AL. Clinical skills education: outcomes of relationships between junior medical students, senior peers and simulated patients. Med Educ. 2008;42(9):901-908. DOI: 10.1111/j.1365-2923.2008.03107.x
6. Young I, Montgomery K, Kearns P, Hayward S, Mellanby E. The benefits of a peer-assisted mock OSCE. Clin Teach. 2014;11(3):214-218. DOI: 10.1111/tct.12112
7. Nomura O, Onishi H, Kato H. Medical students can teach communication skills - a mixed methods study of crossyear peer tutoring. BMC Med Educ. 2017;17(1):103. DOI: 10.1186/s12909-017-0939-7
8. Weyrich P, Celebi N, Schrauth M, Möltner A, Lammerding-Köppel M, Nikendei C. Peer-assisted versus faculty staff-led skills laboratory training: a randomised controlled trial. Med Educ. 2009;43(2):113-120. DOI: 10.1111/j.1365-2923.2008.03252.x
9. Chenot JF, Simmenroth-Nayda A, Koch A, Fischer T, Scherer M, Emmert B, Stanske B, Kochen MM, Himmel W. Can student tutors act as examiners in an objective structured clinical examination? Med Educ. 2007;41(11):1032-1038. DOI: 10.1111/j.1365-2923.2007.02895.x
10. Moineau G, Power B, Pion AJ, Wood TJ, Humphrey-Murto S. Comparison of student examiner to faculty examiner scoring and feedback in an OSCE. Med Educ. 2011;45(2):183-191. DOI: 10.1111/j.1365-2923.2010.03800.x

11. Blank WA, Blankenfeld H, Vogelmann R, Linde K, Schneider A. Can near-peer medical students effectively teach a new curriculum in physical examination? BMC Med Educ. 2013;13:165. DOI: 10.1186/1472-6920-13-165

12. Melcher P, Roth A, Ghanem M, Rotzoll D. Klinisch-praktische Prüfungen in der orthopädischen Lehre: Wer ist der "ideale" Prüfer? Z Orthop Unfall. 2017;155(4):468-475. DOI: 10.1055/s-0043-109022

13. Melcher P, Zajonz D, Roth A, Heyde C, Ghanem M. Peer-assisted teaching student tutors as examiners in an orthopedic surgery OSCE station - pros and cons. GMS Interdiscip Plast Reconstr Surg DGPW. 2016;5:Doc17. DOI: 10.3205/iprs000096

14. Ledig T, Eicher C, Szecsenyi J, Engeser P. AaLplus - ein Anamnese- und Untersuchungskurs für den vorklinischen Studienabschnitt. Z Allg Med. 2014;90(2):76-80.

15. Schwill S, Fahrbach-Veeser J, Moeltner A, Eicher C, Kurczyk S, Pfisterer D, Szecsenyi J, Loukanova S. Peers as OSCE assessors for junior medical students-a review of routine use: a mixed methods study. BMC Med Educ. 2020;20(1):1-12. DOI: 10.1186/s12909-019-1898-y

16. Black P, Wiliam D. Developing the theory of formative assessment. Educ Asse Eval Acc. 2009;21(1):5-31. DOI: 10.1007/s11092-008-9068-5

17. Dolin J, Black P, Harlen W, Andrée Tiberghien A. Exploring Relations Between Formative and Summative Assessment. In: Dolin J, Evans R, editors. Transforming Assessment: Through an interplay between practice, research and policy. Cham, Switzerland: Springer; 2018. p.53-80. DOI: 10.1007/978-3-319-63248-3_3

18. O'Shaughnessy SM, Pauline J. Summative and Formative Assessment in Medicine: The Experience of an Anaesthesia Trainee. Internl J High Educ. 2015;4(2):198-206. DOI: 10.5430/ijhe.v4n2p198

19. Pugh D, Desjardins I, Eva K. How do formative objective structured clinical examinations drive learning? Analysis of residents' perceptions. Med Teach. 2018;40(1):45-52. DOI: 10.1080/0142159X.2017.1388502

20. Lim YS. Students' Perception of Formative Assessment as an Instructional Tool in Medical Education. Med Sci Educ. 2019;29(1):255-263. DOI: 10.1007/s40670-018-00687-w

21. Khan R, Payne MW, Chahine S. Peer assessment in the objective structured clinical examination: A scoping review. Med Teach. 2017;39(7):745-756. DOI: 10.1080/0142159X.2017.1309375

22. Basehore PM, Pomerantz SC, Gentile M. Reliability and benefits of medical student peers in rating complex clinical skills. Med Teach. 2014;36(5):409-414. DOI: 10.3109/0142159X.2014.889287

23. Hochlehnert A, Schultz JH, Möltner A, Timbil S, Brass K, Jünger J. Elektronische Erfassung von Prüfungsleistungen bei OSCE-Prüfungen mit Tablets. GMS Z Med Ausbild. 2015;32(4):Doc41. DOI: 10.3205/zma000983

24. Brennan RL. Generalizability Theory. New York NY: Springer; 2001. DOI: 10.1007/978-1-4757-3456-0

25. Epstein RM. Assessment in Medical Education. N Engl J Med. 2007;356(4):387-396. DOI: 10.1056/NEJMra054784

Corresponding author:
Andreas Möltner
University Heidelberg, Baden-Württemberg Center of Excellence for Assessment in Medicine, Im Neuenheimer Feld 346, D-69120 Heidelberg, Germany
andreas.moeltner@med.uni-heidelberg.de

Please cite as
Möltner A, Lehmann M, Wachter C, Kurczyk S, Schwill S, Loukanova S. Formative assessment of practical skills with peer-assessors: quality features of an OSCE in general medicine at the Heidelberg Medical Faculty. GMS J Med Educ. 2020;37(4):Doc42. DOI: 10.3205/zma001335, URN: urn:nbn:de:0183-zma0013354

This article is freely available from
https://www.egms.de/en/journals/zma/2020-37/zma001335.shtml

Received: *2019-05-14*
Revised: *2020-03-24*
Accepted: *2020-04-15*
Published: *2020-06-15*

Formatives Prüfen praktischer Fertigkeiten mit studentischen Prüfern: Qualitätseigenschaften des OSCE Allgemeinmedizin der Medizinischen Fakultät Heidelberg

Andreas Möltner[1]
Mirijam Lehmann[1]
Cornelia Wachter[2]
Sonia Kurczyk[2]
Simon Schwill[2]
Svetla Loukanova[2]

1 Universität Heidelberg, Kompetenzzentrum für Prüfungen in der Medizin Baden-Württemberg, Heidelberg, Deutschland

2 Universität Heidelberg, Med. Fakultät, Abteilung Allgemeinmedizin und Versorgungsforschung, Heidelberg, Deutschland

Zusammenfassung

Hintergrund: Objective Structured Clinical Examinations (OSCEs) sind mittlerweile ein etabliertes Prüfungsformat an deutschen medizinischen Fakultäten. Üblicherweise werden darin praktische und kommunikative Fertigkeiten von medizinischen Experten summativ bewertet. Der Einsatz des OSCEs als formatives Prüfungsformat mit studentischen Prüfern findet bislang eher wenig Anwendung.
Zielsetzung: Der an der Medizinischen Fakultät Heidelberg im Fach Allgemeinmedizin durchgeführte formative OSCE, der von Peer-Tutoren durchgeführt und bewertet wird, soll hinsichtlich seiner Gütekriterien untersucht und mit denen summativer OSCEs aus anderen Fachbereichen verglichen werden.
Methodik: Schwierigkeiten und Trennschärfen der einzelnen Stationen werden für die summativen sowie den formativen OSCE bestimmt und einander gegenübergestellt. Zur Beurteilung der Messzuverlässigkeit wird eine Analyse der Daten mittels der Generalisierbarkeitstheorie durchgeführt. Zusätzlich findet ein Vergleich zwischen den Bewertungen der studentischen Prüfer und Zweitbewertungen medizinischer Experten statt.
Ergebnisse: Die Stationen des formativen OSCEs weisen ähnliche Schwierigkeiten wie die der summativen Vergleichs-OSCEs auf (P_{form}=0.882; P_{sum}=0.845 – 0.902). Bezüglich der Messzuverlässigkeit zeigen sich keine Unterschiede zwischen dem OSCE Allgemeinmedizin und denen der anderen Fächer. Die Bewertungen der studentischen Prüfer und der medizinischen Experten korrelieren hoch (r=0.888).
Schlussfolgerung: Der formative OSCE Allgemeinmedizin ist hinsichtlich seiner Qualitätskriterien vergleichbar mit denen der summativen Vergleichsformate. Der Einsatz studentischer Prüfer kann bei formativen OSCEs eine verlässliche Alternative zu medizinischen Experten darstellen.

Schlüsselwörter: formativ, OSCE, studentische Prüfer, Generalisierbarkeitstheorie

1. Einleitung

Praktische klinische Fähigkeiten und Anamneseerhebung werden bereits an verschiedenen medizinischen Fakultäten in den vorklinischen Studiensemestern vermittelt und mit Hilfe eines Objektive Structured Clinical Examination (OSCE) überprüft. Es konnte gezeigt werden, dass das frühe Erlernen praktischer Fähigkeiten zu besseren Ergebnissen in klinischen Examensabschnitten und in den klinischen Fertigkeiten führt [1].
Die Vermittlung der Lehrinhalte erfolgt traditionsgemäß durch Lehrärzte der Fakultät, in zunehmendem Maße jedoch auch durch studentische Tutoren höherer Semester. Ein Vorteil von Peer-Tutoren (Peer Assisted Learning, PAL) sind die höhere Akzeptanz durch die Studierenden [2], die niedrigeren Kosten [3], [4] und der Möglichkeit von kleineren Lerngruppen [5]. Zudem profitieren hierbei sowohl die Studierenden durch eine Reduktion von Stress- und Angstfaktoren [6] wie auch die studentischen Tutoren [2], [7] durch die eigene Vertiefung der Lerninhalte. Beim Vergleich der Studententutoren mit Lehrärzten der Fakultät erzielen die Studierenden beim PAL einen gleichen Ergebnisstand in abschließenden Prüfungen [8], [9], [10], [11] und eine gleiche bis höhere Qualität des Feedbacks [10]. Voraussetzungen hierbei sind genau definierte Studententutoren-Schulungen und Checklisten [12], [13]. Seit dem Jahr 2013 werden an der Medizinischen Fakultät Heidelberg im vorklinischen Studienabschnitt im Rahmen der AaLPLUS -Veranstaltungen (AaL: „Anatomie am Lebenden“) der Abteilung Allgemeinmedizin prakti-

sche Fähigkeiten und Anamnesetechniken mit Hilfe von Peer-Tutoren vermittelt und anschließend in einem ebenfalls von Studententutoren durchgeführten formativen OSCE überprüft [14]. Eine detaillierte Darstellung des Programms und der Evaluation des OSCEs durch Studierende und Peer-Tutoren findet sich in [15].
Black und Wiliam [16] sehen fünf wesentliche Aspekte bei formativen Prüfungen. Angepasst an den Kontext der universitären Ausbildung sind dies:

1. Klärung und Austausch von Lernabsichten und Erfolgskriterien
2. Anstoß von effektiven Diskussionen und anderen Lernaufgaben, die das Verständnis der Studierenden für die Lerninhalte belegen
3. Rückmeldungen, die für die Studierenden nützlich sind
4. Aktivierung der Studierenden, als gegenseitige Lernquelle zu fungieren
5. Aktivierung der Studierenden, sich als Initiator ihrer eigenen Lernaktivitäten zu sehen.

Diese Zielsetzungen beinhalten einen ganzen Prozess der Lehre, in dem mehr oder weniger kontinuierlich formative Prüfungen integriert sind. Dies ist in vollem Umfang bei formativen praktischen Prüfungen in Form von OSCEs in der medizinischen Ausbildung logistisch oft schwer zu erfüllen, so dass der hier betrachtete formative OSCE eher als Instrument zu sehen ist [17], der am Ende des vorklinischen Abschnitts des Studiums steht. Um die von Black und Wiliam avisierten Ziele zu erreichen, dürften andere Formen formativer Prüfungsverfahren geeignet sein [18]. Trotz dieser begrenzten Funktion des formativen OSCEs kann erwartet werden, dass er sich positiv auf das Lernverhalten der Prüfungsteilnehmer auswirkt [19], [20].
In einem Übersichtsartikel von Khan et al. aus dem Jahr 2017 werden 13 Publikationen zum Thema „Studierende als Prüfer in OSCEs“ näher dargestellt [21]. Einige der dort aufgeführten Arbeiten untersuchen die Bewertungen von Studierenden und Experten hinsichtlich basaler Kennwerte wie Unterschiede bei den vergebenen Punktzahlen und die Korrelation der Bewertungen durch Studierende und Experten als Prüfer. Eine eingehendere quantitative Analyse, die auch eine Differenzierung von Stations- und Prüfereffekten und deren Konsequenzen für die Messzuverlässigkeit enthält, erfolgt nur in den Arbeiten von Moineau et al. [10] und Basehore et al. [22]. Bei beiden Arbeiten werden Doppelbewertungen an den Stationen durch Studierende und Experten untersucht (in [22] bewerteten die Experten anhand von Videos der Prüfungen). Nicht untersucht wurde jedoch, ob sich studentische Prüfer hinsichtlich des Ausmaßes an Prüfereffekten von Experten unterscheiden.
Neben dem Vergleich von studentischen Prüfern und Experten bei der gleichen formativen Prüfung ist auch die Qualität der formativen Prüfung in Relation zu an der Fakultät etablierten summativen Prüfungen von Interesse. Formative Prüfungen unterscheiden sich von Ihrer Zielsetzung und Struktur (z. B. höhere Bedeutung des Feedbacks) und der Relevanz für die Prüflinge von summativen Prüfungen. Insbesondere Letzteres kann Auswirkungen auf die Messzuverlässigkeit und -genauigkeit haben, z. B. etwa dann, wenn durch eine verringerte Motivation der Prüflinge deren Leistungen weniger differenziert erbracht werden.

Ziel der Studie

Ziel der Studie ist,

1. nachzuweisen, dass Studierende im Kontext formativer Prüfungen praktischer Fertigkeiten, Experten als Prüfer ersetzen können, ohne dass dadurch die Qualität der Prüfung leidet und
2. dass die Qualität solcher formativen Prüfungen die gleichen Standards erreicht wie etablierte summative Prüfungen.

Hierzu soll der im Jahr 2018 durchgeführte formative OSCE Allgemeinmedizin an der Medizinischen Fakultät Heidelberg, bei dem Tutoren als Prüfer eingesetzt werden, hinsichtlich seiner Gütekriterien (Kennwerte der Stationen, Messzuverlässigkeit der Prüfung, Ausmaß von Prüfereffekten) untersucht werden, ein Vergleich mit summativen OSCEs erfolgen und die Übereinstimmung der Bewertungen studentischer Prüfer mit denen von Experten („Supervisoren“) betrachtet werden .
Andere Aspekte des formativen OSCE Allgemeinmedizin mit studentischen Prüfern, wie etwa die Akzeptanz seitens der prüfenden wie auch der geprüften Studierenden, die Einschätzung der Qualität des Feedbacks und des subjektiven Nutzens hinsichtlich der im OSCE abgeprüften Fertigkeiten für die Prüfungsteilnehmer und die Prüfer sind ausführlich in [15] dargestellt. Die vorliegende Studie thematisiert ausschließlich die durch statistische Kennwerte der Prüfungsergebnisse erfassbaren Qualitätseigenschaften des OSCE.
Standardanalysen von Prüfungen umfassen meist basale Kennwerte wie Schwierigkeit, Trennschärfe und Reliabilität (s. 3.1.1). Auf Basis der Generalisierbarkeitstheorie werden darüber hinaus die Facetten (Einflussfaktoren) „Studierende“ (Unterschiede in der Fähigkeit der Studierenden), „Station“ (Unterschied in der Schwierigkeit der Stationen), „Prüfer“ (Unterschied bei der „Strenge“ der Prüfer) und der Interaktion „Station x Prüfer“ (Unterschiedliche Strenge von Prüfern an verschiedenen Stationen) und deren Auswirkungen auf Generalisierbarkeit und absolute Messgenauigkeit (s. 3.1.2) untersucht.
Zum Vergleich der Kennwerte des OSCEs Allgemeinmedizin mit etablierten summativen OSCEs der Medizinischen Fakultät Heidelberg wurden die OSCEs der Fächer Chirurgie und Innere Medizin des Wintersemesters (WS) 2017/2018, des Sommersemesters (SS) 2018 und des WS 2018/2019 herangezogen.
Abschließend erfolgt ein Vergleich von Doppelbewertungen durch studentische Prüfer und Experten innerhalb des formativen OSCEs Allgemeinmedizin (3.2).

2. Methoden

2.1. Durchführung des OSCEs

An dem formativen OSCE Allgemeinmedizin im Mai 2018 nahmen 300 Studierende des vierten Fachsemesters teil. Der OSCE fand an zwei Tagen statt und umfasste vier Stationen. Eine der vier Stationen („Venöse Blutentnahme“) wurde von allen Studierenden durchlaufen. An zwei Stationen mussten verschiedene klinische Untersuchungen durchgeführt werden. Diese Stationen waren für die teilnehmenden Studierenden nicht identisch, sondern wechselten zwischen den verschiedenen Parcours. Insgesamt wurden 11 verschiedene Aufgaben (Allgemeine Untersuchung des Abdomens, Untersuchung von Milz/Niere/Appendizitiszeichen, Blutdruckmessung, Untersuchung des Herzens, der Leber, des Lymphknotenstatus, des Pulsstatus, der Schilddrüse, des Thorax, der Wirbelsäule und eine neurologische Untersuchung) verwendet. Weiter musste eine vollständige Anamnese durchgeführt werden. Auch hier wechselten die Inhalte (Rücken-, Bauch- und Kopfschmerz). Für die klinischen Untersuchungen und die Anamnesen wurden geschulte Schauspielpatienten eingesetzt. Die Inhalte der Stationen und die wesentlichen Kriterien zur Beurteilung waren den teilnehmenden Studierenden aus den Kursen und -materialien bekannt.
Jeder Teilnehmende durchlief insgesamt vier Stationen von achtminütiger Dauer (5 Minuten pro Aufgabe und 3 Minuten Feedback). Die Bewertung der Leistung erfolgte anhand von Checklisten durch basisdidaktisch geschulte Studierende, die mindestens im sechsten Semester waren. Insgesamt konnten an den Stationen jeweils 25 Punkte erreicht werden. Eine Ausnahme hiervon bildeten die drei Stationen, an denen eine Anamnese durchgeführt werden musste. Bei diesen waren 30 Punkte zu erreichen. Als Prüfer waren 32 Studierende im Einsatz, von denen im Verlauf des OSCEs 26 an mehreren (bis zu fünf) Stationen geprüft haben (siehe Tabelle 1). Die Erfassung der Bewertungen erfolgte mit Tablets (Programm tOSCE des UCAN-Prüfungsverbunds) [23].
Zur Qualitätskontrolle der Durchführung und Bewertung waren fünf Supervisoren eingesetzt, die stichprobenartig Zweitbewertungen durchführten (insgesamt 135 Bewertungen). Die geschulten Prüfer waren (ärztliche) Mitarbeiter der Abteilung Allgemeinmedizin und für die Beurteilung der kommunikativen Fertigkeiten an den drei Anamnesestationen Lehrende der Abteilung Medizinische Psychologie.

2.2. Vergleich mit summativen OSCEs

Zum Vergleich der Gütekriterien des OSCEs Allgemeinmedizin wurden sechs OSCEs der Fächer Chirurgie und Innere Medizin der Wintersemester 2017/2018 und 2018/2019 und des Sommersemesters 2018 der Medizinischen Fakultät Heidelberg herangezogen. Durch die Einbeziehung mehrerer Vergleichs-OSCEs aus zwei verschiedenen Fächern und Semestern wird sichergestellt, dass bei den Vergleichs-OSCEs eine Abschätzung der Variabilität ihrer Kennwerte (z. B. Anteil von Prüfereinflüssen) vorgenommen werden kann.
Die OSCEs der Inneren Medizin umfassten jeweils 10, die der Chirurgie 13 Stationen. An allen Stationen dieser OSCEs konnten maximal 25 Punkte erreicht werden (siehe Tabelle 2). Diese OSCEs wurden jeweils an zwei bis drei Tagen in jeweils zwei zeitlich parallelen Parcours durchgeführt. Die Stationen wurden teilweise in den verschiedenen Parcours gewechselt. Die beiden Fächer Innere Medizin und Chirurgie wurden gewählt, da bei diesen

1. an denselben Stationen unterschiedliche Prüfer und
2. die Prüfer i. A. an verschiedenen Stationen eingesetzt wurden.

Dies ermöglicht bei der Auswertung eine Abschätzung von Prüfer-, Stations- und dem Interaktionseffekt Station x Station.

2.3. Statistische Analyse

Für die Stationen aller genannten OSCEs wurden Schwierigkeiten P und korrigierte Trennschärfen r_{it} (Korrelationen der an einer Station erreichten Punktzahl mit den an allen anderen Stationen erreichten Punkten) sowie die gemittelten Interkorrelationen mit allen anderen Stationen r_{ij} (Average inter-item correlation) bestimmt. Als Korrelationsmaß wurde durchweg die Produkt-Moment-Korrelation (nach Pearson) verwandt.
Um eine Gleichwertigkeit der Stationen zu erzielen, wurden für alle dargestellten Analysen die an den Anamnesestationen, an denen 30 Punkte zu erreichen waren, erzielten Punktwerte auf den Bereich von 0-25 Punkten reskaliert.
Zur Beurteilung der Messzuverlässigkeit wurde eine Analyse der Daten mittels der Generalisierbarkeitstheorie [24] durchgeführt. Die betrachteten Facetten waren „Studierende“, „Stationen“, „Prüfer“ und die Interaktion „Station x Prüfer“. Aus den durch die Anwendung der Generalisierbarkeitstheorie gefundenen Varianzkomponenten lassen sich die „Generalizability“ $E\rho^2$ (als Analogon zur internen Konsistenz/Cronbachs α) und die „Dependability“ Φ als Maß der absoluten Messgenauigkeit bestimmen:
Bezeichne *n* die Zahl der Stationen, so ist

$$E\rho^2 = \frac{\sigma^2_{Stud}}{\sigma^2_{Stud} + \sigma^2_{Resid}/n}$$

$$\theta = \frac{\sigma^2_{Stud}}{\sigma^2_{Stud} + (\sigma^2_{Station} + \sigma^2_{Prüfer} + \sigma^2_{Station*Prüfer} + \sigma^2_{Resid})/n}$$

Zur Analyse der Übereinstimmung der Bewertungen der studentischen Prüfer und der Supervisoren wurden je Station die vergebenen Punktzahlen verglichen (Wilcoxon-Vorzeichen-Rang-Tests) und die Korrelationen bestimmt. Weiterhin erfolgte eine Varianzanalyse des Gesamtdatensatzes (Prüfer und Supervisoren) mit dem festen Faktor „Studentischer Prüfer/Supervisor“ und den Facetten

Tabelle 1: Anzahl der Bewertungen im OSCE Allgemeinmedizin 2018 nach Stationen.

Station	Bewertungen	Supervisionen
Anamnese Bauch	66	7
Anamnese Kopf	94	8
Anamnese Rücken	140	22
KU Abdomen I: Allgemeine Untersuchung des Abdomens	73	9
KU Abdomen II: Milz, Niere und Appendizitiszeichen	47	0
KU Blutdruckmessung	46	0
KU Herz	47	12
KU Leber	24	8
KU Lymphknotenstatus	60	8
KU Neurologie	60	24
KU Pulsstatus	71	8
KU Schilddrüse	50	8
KU Thorax/Lunge	50	8
KU Wirbelsäule	72	6
Venöse Blutentnahme	300	7
	1200	135

Tabelle 2: Anzahl der Teilnehmer, Stationen und Prüfer im OSCE Allgemeinmedizin und den OSCEs der Chirurgie und Inneren Medizin WS2017/2018 bis 2018/2019.

OSCE	Teilnehmer	Stationen	Prüfer
Allgemeinmedizin 2018	300	15	32
Innere Medizin WS18	160	28	24
Chirurgie WS18	173	52	58
Innere SS18	193	28	26
Chirurgie SS18	182	52	56
Innere WS17	179	27	30
Chirurgie WS17	145	47	63

„Studierende“, „Stationen“, „studentische Prüfer“, „Supervisor“ sowie der Interaktion „Station x Prüfer“.
Anmerkung: Bei der Analyse mittels der Generalisierbarkeitstheorie muss unterschieden werden zwischen sog. festen und Zufallsfaktoren („fixed“ bzw. „random factors“). Wird die Facette „Student“ als Zufallsfaktor betrachtet, so intendiert man eine Verallgemeinerbarkeit auf äquivalente Studentengruppen (also im selben Semester, gleiche demographische Zusammensetzung, gleichwertige Lehre etc.). Die in der untersuchten Prüfung betrachtete Studierendengruppe ist demzufolge als Stichprobe aus einer Population aufzufassen. Ähnliches gilt für die Facette „Station“: Als Zufallsfaktor steht die Verallgemeinerbarkeit auf äquivalent konstruierte Stationen im Zentrum, bei der Facette „Prüfer“ die Einbeziehung von Prüfern aus einer potentiellen Gruppe von Prüfern. Bei der Modellierung von Station oder Prüfer als fester Faktor zielt man hingegen auf die in der Prüfung tatsächlich eingesetzten Stationen bzw. Prüfer ab: Sind einzelnen Stationen besonders leicht oder schwer, sind Prüfer zu streng oder zu nachsichtig? Da in der vorliegenden Studie die Verallgemeinerbarkeit im Fokus steht, werden nur die Ergebnisse für die Analysen mit „Student“, „Station“ und „Prüfer“ als Zufallsfaktoren dargestellt.
Die statistischen Analysen wurden mit R Version 3.5.1 durchgeführt. Für die Mixed-Model-Analysen zur Auswertung mit dem Modell der Generalisierbarkeitstheorie wurden die Pakete „lme4“ und „lmerTest“ verwendet.

3. Ergebnisse

3.1. Kennwerte der Prüfung

3.1.1. Schwierigkeiten und Trennschärfe der Stationen

Die Basiskennwerte (mittlere erreichte Punktzahl x, Schwierigkeit P und korrigierte Trennschärfe r_{it}) der an den Stationen erzielten Punktwerte sind in Tabelle 3 aufgeführt. Abbildung 1 enthält eine grafische Darstellung der Verteilungen als Boxplot.
Die Schwierigkeiten an den einzelnen Stationen reichen von P=0.794 bei der Station „Anamnese Bauch“ bis P=0.959 an der Station „KU Blutdruck“. Im Mittel wurden 87.632 von maximal 100 Punkten erreicht. Man beachte, dass im Unterschied zu dichotomen Items, bei denen nur 0 oder 1 Punkt erreicht werden kann, bei feiner granulierten Bewertungen (hier 0-25 Punkte) Trennschärfen u. U. auch dann interpretiert werden können, wenn die Schwierigkeiten numerisch hoch sind.
Elf der 15 Stationen weisen Part-whole-korrigierte Trennschärfen von über 0.300 auf, zwei Stationen liegen mit Trennschärfen von 0.276 und 0.296 knapp unter dieser Grenze („KU Blutdruck“ bzw. „KU Neurologie“). Deutlich niedriger sind die der Stationen „KU Leber“ mit r_{it}=0.112 und „Pulsstatus“ mit r_{it}=0.099.

Tabelle 3: Kennwerte der Stationen des formativen OSCE Allgemeinmedizin 2018.

Station	*N*	$\bar{x}$	*P*	r_{it}
Anamnese Bauch[1]	66	19.861	0.794	0.384
Anamnese Kopf[1]	94	20.372	0.815	0.392
Anamnese Rücken[1]	140	20.577	0.823	0.357
KU Abdomen I	73	21.863	0.875	0.541
KU Abdomen II	47	23.064	0.923	0.663
KU Blutdruck	46	23.978	0.959	0.276
KU Herz	47	22.234	0.889	0.512
KU Leber	24	22.458	0.898	0.112
KU Lymphknoten	60	22.767	0.911	0.451
KU Neurologie	60	21.783	0.871	0.296
KU Pulsstatus	71	22.944	0.918	0.099
KU Schilddrüse	50	22.800	0.912	0.407
KU Thorax	50	21.180	0.847	0.485
KU Wirbelsäule	72	22.361	0.894	0.419
Venöse Blutentnahme	300	22.350	0.894	0.399

[1]Auf den Bereich von 0-25 reskalierte Punktwerte

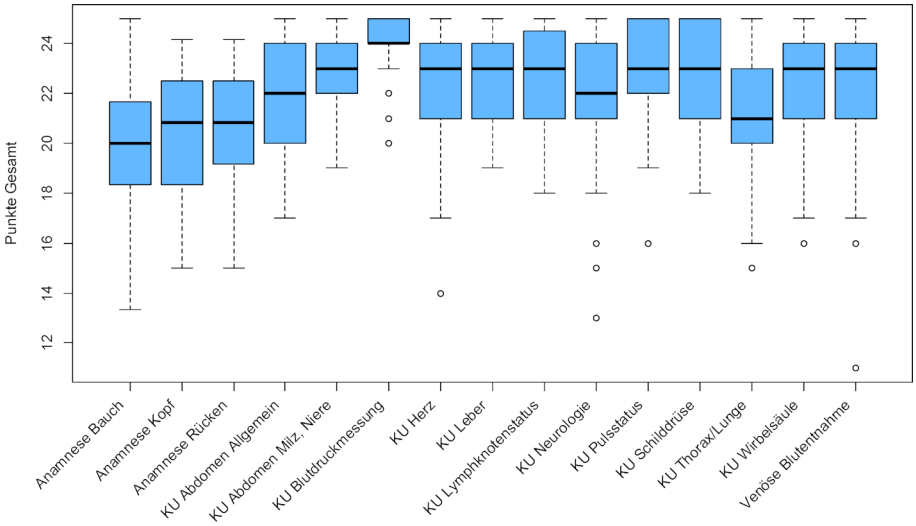

Abbildung 1: Verteilungen der erreichten Punktzahlen an den Stationen des formativen OSCE Allgemeinmedizin. Die Station „Vollständige Anamnese", an der im Original-OSCE 30 Punkte erreicht werden konnten, ist auf den Bereich von 0-25 Punkte reskaliert.

Vergleich mit summativen OSCEs

Abbildung 2 zeigt die Verteilung der an den Stationen erreichten Punktzahlen des OSCEs Allgemeinmedizin im Vergleich zu den summativen OSCEs der Inneren Medizin und der Chirurgie der vergangenen drei Semester (siehe auch Tabelle 4).

Im Vergleich zu den betrachteten OSCEs der Inneren Medizin und der Chirurgie waren die Stationen des OSCE Allgemeinmedizin annähernd gleich schwer (P=0.882 gegenüber P=0.876).

Die korrigierten Trennschärfen waren im Mittel etwas geringer als bei den Vergleichs-OSCEs, lediglich der OSCE Innere Medizin SS 2018 wies hier niedrigere Werte auf (r_{it}=0.358 gegenüber 0.386, siehe Tabelle 4 und Abbildung 3). Bei diesem Vergleich ist jedoch zu berücksichtigen, dass beim OSCE Allgemeinmedizin die für die korrigierte Trennschärfe verwendete Punktsumme der anderen Stationen nur aus drei Stationen bestimmt wird, im Gegensatz zur Inneren Medizin und der Chirurgie mit neun bzw. zwölf Stationen. Damit ist diese Summe beim OSCE Allgemeinmedizin mit mehr Fehlervarianz behaftet. Eine bessere Vergleichsmöglichkeit bietet hier das Mittel aus allen Korrelationen der Punktsumme aus einer Station mit allen anderen Stationen rij („averaged inter-item correlation"). Hier zeigt sich, dass jeweils drei der Vergleichs-OSCEs niedrigere und höhere Werte aufweisen (siehe Tabelle 4 und Abbildung 4).

3.1.2. Messzuverlässigkeit

Zur Analyse der Messzuverlässigkeit wurden Verfahren der Generalisierbarkeitstheorie eingesetzt. Analysiert wurde ein Modell mit den Facetten „Studierender", „Station", „Prüfer" und der Interaktion „Station x Prüfer". In Tabelle 5 sind die geschätzten Varianzkomponenten der Facetten aufgeführt.

Nahezu 53% der Varianz können durch die Effekte des Modells erklärt werden, wobei 22% auf die Unterschiede

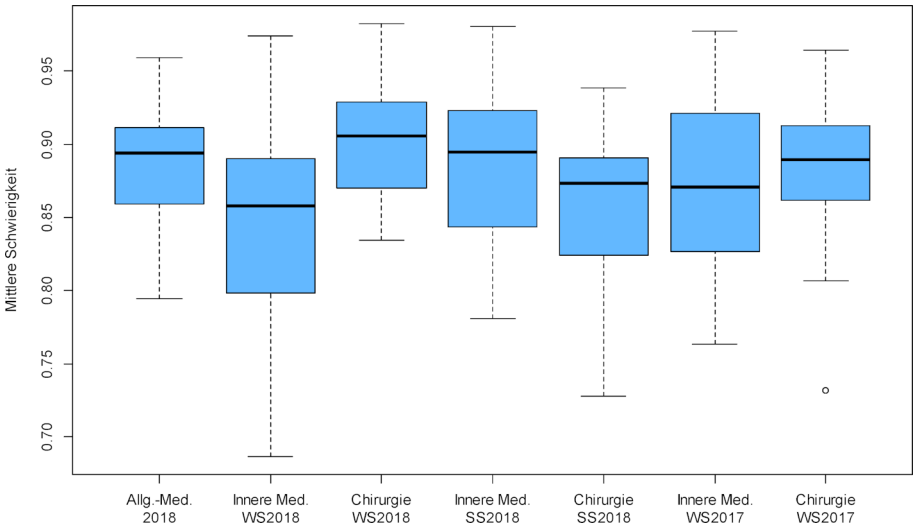

Abbildung 2: Verteilung der mittleren erreichten Schwierigkeiten P an den Stationen des formativen OSCE Allgemeinmedizin 2018 und der summativen OSCEs Innere Medizin und Chirurgie Wintersemester 2017/18 bis Wintersemester 2018/2019.

Tabelle 4: Mittlere Schwierigkeiten, Trennschärfen und Interkorrelationen mit anderen Stationen des OSCE Allgemeinmedizin und der OSCEs Innere Medizin und Chirurgie des WS 2017 – WS 2018.

OSCE	P	r_{it}	r_{ij}
Allgemeinmedizin 2018	0.882	0.386	0.285
Innere WS18	0.845	0.535	0.359
Chirurgie WS18	0.902	0.417	0.233
Innere SS18	0.884	0.358	0.188
Chirurgie SS18	0.858	0.553	0.356
Innere WS17	0.871	0.487	0.292
Chirurgie WS17	0.883	0.461	0.266

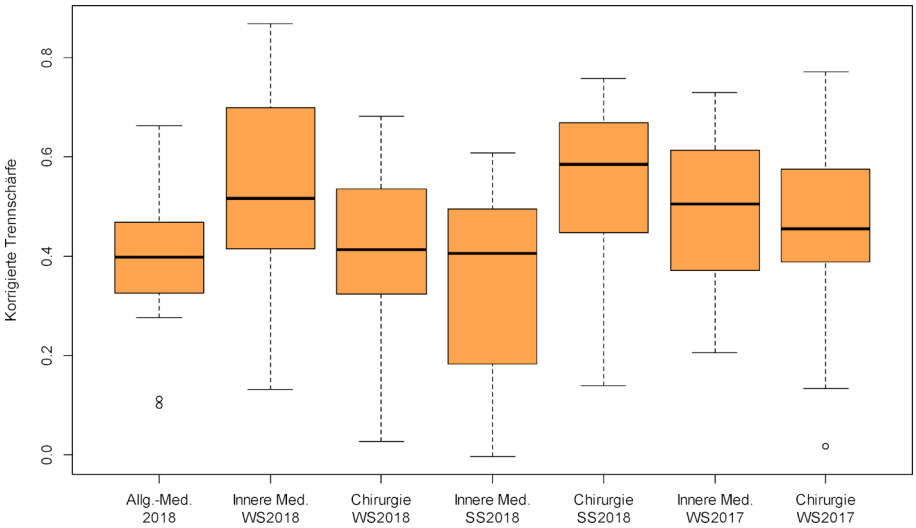

Abbildung 3: Verteilung der korrigierten Trennschärfen rit an den Stationen des formativen OSCE Allgemeinmedizin 2018 und der summativen OSCEs Innere Medizin und Chirurgie Wintersemester 2017/18 bis Wintersemester 2018/2019.

zwischen den Studierenden hinsichtlich ihrer Leistungen zurückgeführt werden können. Auf die Variabilität der Stationen entfallen 21%, die zusammengefassten Prüfereinflüsse betragen etwa 10%. Dabei ist der Interaktionseffekt Station x Prüfer nicht als signifikant von 0 verschieden nachweisbar.

Die zu erwartende Korrelation der beim OSCE erreichten Punktwerte mit einem äquivalenten OSCE beträgt $E\rho^2$=0.647. In diesen Werte gehen die Effekte von Station und Prüfer nicht mit ein, da bei einem äquivalenten Parcours alle Studierenden die gleichen Stationen mit den gleichen Prüfern durchlaufen, ihre erreichte Punktsumme daher durch diese Facetten nur durch einen für alle konstanten Wert verändert sind, der bei einer Korrelation nicht berücksichtigt wird ($E\rho^2$ ist somit ein Maß für die relative Messgenauigkeit). Im Unterschied dazu berücksichtigt die Dependability Φ als Maß für die absolute

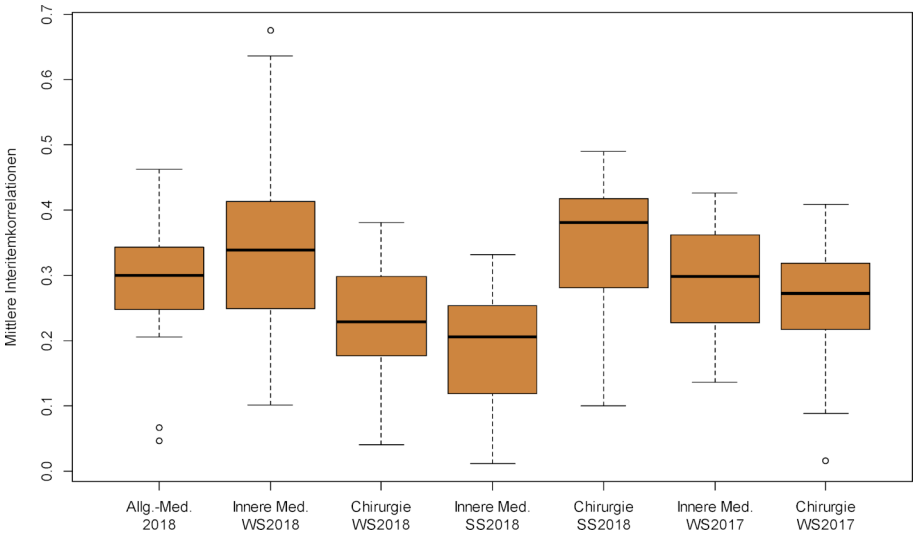

Abbildung 4: Verteilung des mittleren Interitemkorrelationen r_{ij} (Korrelationen der an einer Station erreichten Punktzahl mit den jeweils anderen Stationen) des formativen OSCE Allgemeinmedizin 2018 und der summativen OSCEs Innere Medizin und Chirurgie Wintersemester 2017/18 bis Wintersemester 2018/2019.

Tabelle 5: Varianzkomponenten für die Facetten Studierender, Station, Prüfer und Station x Prüfer. Die Standardabweichung gibt die Größe des Einflusses des jeweiligen Effekts in Punkten an einer Station an.

Facette	Varianz	Std.-Abw	Varianz [%]	*p*
Studierender	1.293	1.137	21.64	<0.001
Station	1.253	1.119	20.96	<0.001
Prüfer	0.544	0.738	9.10	<0.001
Station x Prüfer	0.060	0.246	1.01	0.337
Residuum	2.825	1.681	47.28	

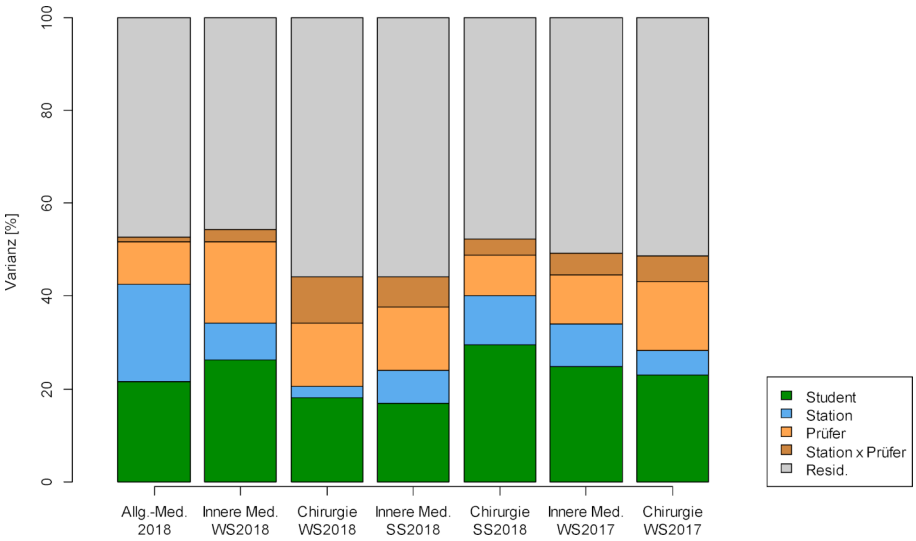

Abbildung 5: Prozentuale Aufteilung der Varianz des OSCE Allgemeinmedizin und der OSCEs Innere Medizin und Chirurgie von WS 2017 bis WS 2018. Die Gesamtvarianz unterteilt sich in die Komponenten „Studierender", „Station", „Prüfer", der Interaktion „Prüfer x Station" und der Residualvarianz.

Messgenauigkeit diese Faktoren, und beträgt für die Prüfung Φ=0.525.

Vergleich mit summativen OSCEs

Abbildung 5 zeigt graphisch die prozentualen Anteile der Varianzkomponenten für die OSCEs. Ein Qualitätsvergleich des OSCEs Allgemeinmedizin mit denen der Inneren Medizin und der Chirurgie hinsichtlich der Stationsqualität und des Umfangs der Prüfereinflüsse muss die unterschiedliche Zahl von Stationen berücksichtigen. Bei Normierung auf einem Parcours von zehn Stationen erhält man die in Tabelle 6 aufgeführten Werte. Es zeigt sich, dass für $E\rho^2$ drei der sechs Vergleichs-OSCEs niedrigere wie auch höhere Werte aufweisen. Die absolute Genauigkeit ist bei vier Vergleichs-OSCEs höher. Wie aus Abbil-

Tabelle 6: Geschätzte Generalizability $E\rho^2$ und Dependability Φ für den OSCE Allgemeinmedizin und die OSCEs Innere Medizin und Chirurgie bei Annahme eines Parcours von 10 Stationen.

OSCE	$E\rho^2$	Φ
Allgemeinmedizin 2018	0.821	0.734
Innere WS18	0.851	0.780
Chirurgie WS18	0.765	0.689
Innere SS18	0.751	0.668
Chirurgie SS18	0.861	0.807
Innere WS17	0.830	0.767
Chirurgie WS17	0.818	0.750

Tabelle 7: Vergleich der Bewertungen durch Prüfer und Supervisoren: Mittlere Punktzahlen der Prüfer (x), mittlere Punktzahlen der Supervisoren (xS), Signifikanzwert des Testes auf Unterschied (Vorzeichen-Rang-Test von Wilcoxon, p) und Korrelation der Bewertungen (r).

Station	n	$\bar{x}$	$\bar{x}_S$	p	r
Anamnese Bauch	7	20.595	19.643	0.034	0.989
Anamnese Kopf	8	18.854	18.229	0.389	0.926
Anamnese Rücken	22	18.750	18.902	0.483	0.729
KU Abdomen: Allgemein	9	22.667	22.889	0.586	0.841
KU Herz	12	21.083	21.500	0.120	0.946
KU Leber	8	21.500	21.375	0.773	0.917
KU Lymphknotenstatus	8	21.500	22.000	0.265	0.783
KU Neurologie	24	22.250	22.292	0.903	0.904
KU Pulsstatus	8	23.875	24.250	0.149	0.787
KU Schilddrüse	8	21.500	22.125	0.269	0.851
KU Thorax/Lunge	8	20.500	20.125	0.345	0.965
KU Wirbelsäule	6	21.500	22.000	0.149	0.986
Venöse Blutentnahme	7	21.857	21.857	1.000	0.917

Tabelle 8: Varianzkomponenten der Analyse der Beurteilungen von studentischen Prüfern und Supervisoren.

Facette	Varianz	Std.-Abw	p
Student	1.392	1.180	<0.001
Station	1.266	1.125	<0.001
Tutor (Prüfer)	0.490	0.700	<0.001
Supervisor (Prüfer)	0.311	0.557	0.117
Station x Prüfer	0.106	0.325	0.079
Residuum	2.716	1.648	

dung 5 zu entnehmen ist, ist dies im Wesentlichen auf die höhere Variabilität der Stationen zurückzuführen.

3.2. Supervision

Bei 135 Bewertungen wurde eine Zweitbewertung durch einen Supervisor (ärztliche Mitarbeiter der Abteilung Allgemeinmedizin und Medizinische Psychologie) vorgenommen, die der Qualitätssicherung des OSCE dient (siehe Tabelle 1). In Tabelle 7 sind die Mittelwerte der Bewertungen durch die Prüfer sowie die der Supervisoren für die Stationen mit Doppelbewertungen aufgeführt, zusätzlich ist der Signifikanzwert des Tests auf Unterschied der Bewertungen (Wilcoxon-Vorzeichen-Rang-Test) angegeben. Nur bei einer Station („Anamnese Bauch“) zeigt sich ein statistisch signifikanter Unterschied.
Tabelle 7 enthält weiterhin die Korrelationen zwischen Prüfern und Supervisoren an den Stationen, diese lagen zwischen 0.729 und 0.989. Als Beispiele sind die Streudiagramme (Blasendiagramme) der Bewertungen für die Stationen „Anamnese Rücken“ und „KU Neurologie“ in Abbildung 6 dargestellt.
Eine Gesamtanalyse auf Basis der Generalisierbarkeitstheorie aller Daten (studentische Prüfer und Supervisoren) mit der Prüfergruppe als fester Faktor und mit getrennten Varianzkomponenten für die beiden Prüfergruppen enthält Tabelle 8. Die Supervisoren vergeben 0.568 Punkte weniger als die studentischen Prüfer, der Effekt ist jedoch nicht signifikant (p=0.152). Die Prüfereffekte haben eine Standardabweichung von 0.700 Punkten (vgl. auch Tabelle 5). Bei den fünf Supervisoren kann keine von Null verschiedene Varianzkomponente nachgewiesen werden (p=0.117), was gleichbedeutend damit ist, dass kein Unterschied hinsichtlich ihrer Strenge nachzuweisen ist.

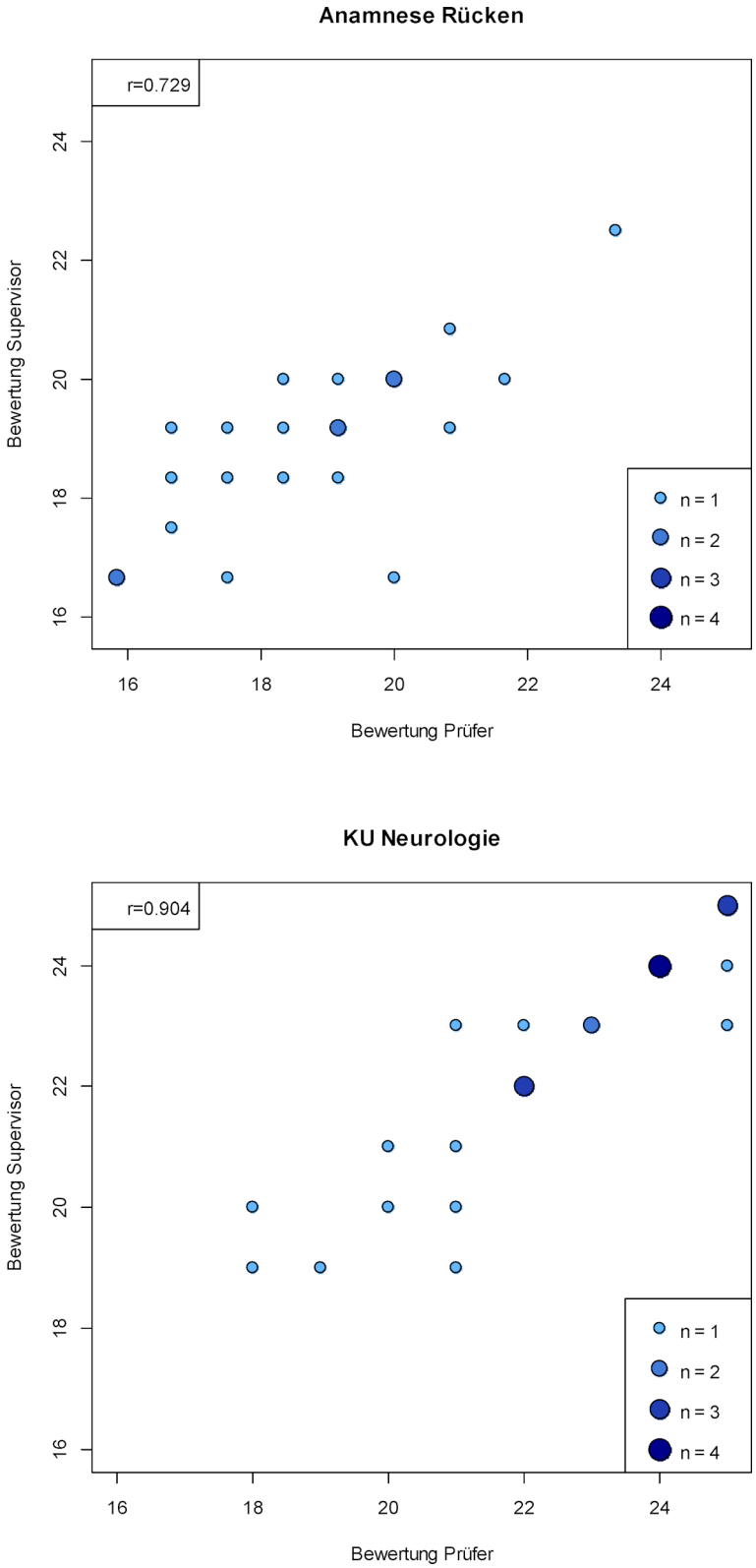

Abbildung 6: Streudiagramme (Blasendiagramme) der Bewertungen durch Prüfer und Supervisoren an den Stationen „Vollständige Anamnese Rücken“ und „KU Neurologie“ (die Kreisgröße repräsentiert die Anzahl mehrfach vorhandener Datenpunkte mit gleichen Werten).

4. Diskussion

Die Ergebnisse zeigen, dass die Stationen des OSCE Allgemeinmedizin 2018 im Wesentlichen die gleichen Qualitätskriterien erfüllen wie die Stationen, die in den seit Jahren etablierten OSCEs der Fächer Chirurgie und Innere Medizin geprüft werden. Bei zwei der klinischen Untersuchungsstationen ist eine Überprüfung auf Grund niedriger Trennschärfen angeraten. Die Übereinstimmung der Bewertungen der studentischen Prüfer mit denen der Supervisoren kann an allen Stationen als gut bis sehr gut bezeichnet werden, systematische Unterschiede zwischen den Bewertungen der studentischen Prüfer und den Supervisoren sind nicht nachzuweisen. Ein relativer Einfluss der Prüfer ist zwar vorhanden, die Prüfereffekte sind tendenziell sogar niedriger als bei den Vergleichs-OSCEs. Die auf zehn Stationen normierte Generalisierbarkeit liegt im OSCE Allgemeinmedizin mit $E\rho^2$=0.82 gegenüber den beiden im Review von Khan [21] genannten Arbeiten, in denen eine Analyse mittels der Generalisierbarkeitstheorie erfolgte, in [10] merklich, in [22] marginal höher ($E\rho^2$=0.51 für die Checkliste und $E\rho^2$=0.63 für den „global score“ bzw. $E\rho^2$=0.80 für den „total score“).
Sofern man von der Anzahl der Stationen absieht, ist die Messzuverlässigkeit der OSCE-Prüfung Allgemeinmedizin vollständig im Rahmen der summativen Vergleichs-OSCEs der Fächer Chirurgie und Innere Medizin der letzten drei Semester.
Damit ist gezeigt, dass bei entsprechender Vorbereitung

1. Studierende statt Experten als Prüfer praktischer Fertigkeiten eingesetzt werden können und
2. die Qualität einer formativen Prüfung mit studentischen Prüfern ähnlich hoch ist wie die etablierter summativer OSCEs mit Experten als Prüfern.

Da die Durchführung praktischer formativer Prüfungen, die den Kenntnisstand für die Studierenden selbst wie auch für Lehrende strukturiert erfassen, an den Fakultäten häufig an der Verfügbarkeit von Prüfern des Lehrkörpers scheitert, können Studierende höherer Fachsemester hier einen vollwertigen Ersatz bieten.
Einzige Schwäche des OSCE Allgemeinmedizin ist die geringe Zahl von vier Stationen, die die Prüfungsteilnehmerinnen und -teilnehmer zu durchlaufen haben. Die Tatsache, dass mit vier Stationen keine Messzuverlässigkeit zu erreichen ist, die den Anforderungen an qualitativ hochwertige Prüfungen genügt, ist jedoch wenig überraschend. Sie steht im Einklang mit der Literatur, in der für OSCEs deutlich höhere Stationszahlen gefordert werden, um als aussagekräftig einzustufende Gesamtbewertungen zu erhalten [25].
Die Analyse anderer formativer Prüfungen, in denen Studierende als Prüfer fungieren, ist natürlich wünschenswert, da aus dem hier vorgestellten Einzelfall keine Verallgemeinerung auf andere Institutionen, Rahmenbedingungen o. ä. möglich ist. Solche Untersuchungen könnten zeigen, welche Voraussetzungen für den Einsatz studentischer Prüfer gegeben sein müssen, um teststatistisch zufriedenstellende und aussagekräftige Leistungsbeurteilungen zu gewinnen. Limitationen: Die stichprobenartigen Zweitbewertungen durch die Supervisoren wurden nicht systematisch durchgeführt, so dass die Vergleiche mit den studentischen Bewertern teils auf sehr geringen Datenzahlen beruhen (siehe Tabelle 7). Ebenfalls verbesserungswürdig ist die Systematik der Zuordnung der beiden klinischen Untersuchungsstationen aus der Menge der elf verfügbaren Stationen zu den Prüfungsteilnehmerinnen und -teilnehmern.

5. Schlussfolgerung

Insgesamt zeigt der OSCE Allgemeinmedizin, dass es möglich ist, mit studentischen Prüfern eine große Zahl an Studierenden zu beurteilen und damit qualitativ hochwertige formative praktische Prüfungen durchzuführen. Die Einbindung von Studierenden in den Prozess der Erstellung formativer Leistungsbeurteilungen stellt damit eine für die medizinischen Fakultäten praktikable Möglichkeit dar, die allseits anerkannten Vorteile von Feedback in der Hochschullehre mit Hilfe strukturierter Leistungserfassungen zu nutzen.

Förderung

Die Arbeit entstand im Rahmen des vom Bundesministerium für Bildung und Forschung geförderten Projekts MERLIN II (01PL17011C).

Interessenkonflikt

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

Literatur

1. Swierszcz J, Stalmach-Przygoda A, Kuzma M, Jablonski K, Cegielny T, Skrzypek A, Wieczorek-Surdacka E, Kruszelnicka O, Chmura K, Chyrchel B, Surdacki A, Nowakowski M. How does preclinical laboratory training impact physical examination skills during the first clinical year? A retrospective analysis of routinely collected objective structured clinical examination scores among the first two matriculating classes of a reformed curriculum in one Polish medical school. BMJ Open. 2017;7(8):e017748. DOI: 10.1136/bmjopen-2017-017748
2. Khalid H, Shahid S, Punjabi N, Sahdev N. An integrated 2-year clinical skills peer tutoring scheme in a UK-based medical school: perceptions of tutees and peer tutors. Adv Med Educ Pract. 2018;9:423-432. DOI: 10.2147/AMEP.S159502
3. Bosse HM, Nickel M, Huwendiek S, Schultz JH, Nikendei C. Cost-effectiveness of peer role play and standardized patients in undergraduate communication training. BMC Med Educ. 2015;15:138. DOI: 10.1186/s12909-015-0468-1
4. Lee CB, Madrazo L, Khan U, Thangarasa T, McConnell M, Khamisa K. A student-initiated objective structured clinical examination as a sustainable cost-effective learning experience. Med Educ Online. 2018;23(1):1440111. DOI: 10.1080/10872981.2018.1440111

5. Hudson JN, Tonkin AL. Clinical skills education: outcomes of relationships between junior medical students, senior peers and simulated patients. Med Educ. 2008;42(9):901-908. DOI: 10.1111/j.1365-2923.2008.03107.x
6. Young I, Montgomery K, Kearns P, Hayward S, Mellanby E. The benefits of a peer-assisted mock OSCE. Clin Teach. 2014;11(3):214-218. DOI: 10.1111/tct.12112
7. Nomura O, Onishi H, Kato H. Medical students can teach communication skills - a mixed methods study of crossyear peer tutoring. BMC Med Educ. 2017;17(1):103. DOI: 10.1186/s12909-017-0939-7
8. Weyrich P, Celebi N, Schrauth M, Möltner A, Lammerding-Köppel M, Nikendei C. Peer-assisted versus faculty staff-led skills laboratory training: a randomised controlled trial. Med Educ. 2009;43(2):113-120. DOI: 10.1111/j.1365-2923.2008.03252.x
9. Chenot JF, Simmenroth-Nayda A, Koch A, Fischer T, Scherer M, Emmert B, Stanske B, Kochen MM, Himmel W. Can student tutors act as examiners in an objective structured clinical examination? Med Educ. 2007;41(11):1032-1038. DOI: 10.1111/j.1365-2923.2007.02895.x
10. Moineau G, Power B, Pion AJ, Wood TJ, Humphrey-Murto S. Comparison of student examiner to faculty examiner scoring and feedback in an OSCE. Med Educ. 2011;45(2):183-191. DOI: 10.1111/j.1365-2923.2010.03800.x
11. Blank WA, Blankenfeld H, Vogelmann R, Linde K, Schneider A. Can near-peer medical students effectively teach a new curriculum in physical examination? BMC Med Educ. 2013;13:165. DOI: 10.1186/1472-6920-13-165
12. Melcher P, Roth A, Ghanem M, Rotzoll D. Klinisch-praktische Prüfungen in der orthopädischen Lehre: Wer ist der "ideale" Prüfer? Z Orthop Unfall. 2017;155(4):468-475. DOI: 10.1055/s-0043-109022
13. Melcher P, Zajonz D, Roth A, Heyde C, Ghanem M. Peer-assisted teaching student tutors as examiners in an orthopedic surgery OSCE station - pros and cons. GMS Interdiscip Plast Reconstr Surg DGPW. 2016;5:Doc17. DOI: 10.3205/iprs000096
14. Ledig T, Eicher C, Szecsenyi J, Engeser P. AaLplus - ein Anamnese- und Untersuchungskurs für den vorklinischen Studienabschnitt. Z Allg Med. 2014;90(2):76-80.
15. Schwill S, Fahrbach-Veeser J, Moeltner A, Eicher C, Kurczyk S, Pfisterer D, Szecsenyi J, Loukanova S. Peers as OSCE assessors for junior medical students-a review of routine use: a mixed methods study. BMC Med Educ. 2020;20(1):1-12. DOI: 10.1186/s12909-019-1898-y
16. Black P, Wiliam D. Developing the theory of formative assessment. Educ Asse Eval Acc. 2009;21(1):5-31. DOI: 10.1007/s11092-008-9068-5
17. Dolin J, Black P, Harlen W, Andrée Tiberghien A. Exploring Relations Between Formative and Summative Assessment. In: Dolin J, Evans R, editors. Transforming Assessment: Through an interplay between practice, research and policy. Cham, Switzerland: Springer; 2018. p.53-80. DOI: 10.1007/978-3-319-63248-3_3
18. O'Shaughnessy SM, Pauline J. Summative and Formative Assessment in Medicine: The Experience of an Anaesthesia Trainee. Internl J High Educ. 2015;4(2):198-206. DOI: 10.5430/ijhe.v4n2p198
19. Pugh D, Desjardins I, Eva K. How do formative objective structured clinical examinations drive learning? Analysis of residents' perceptions. Med Teach. 2018;40(1):45-52. DOI: 10.1080/0142159X.2017.1388502
20. Lim YS. Students' Perception of Formative Assessment as an Instructional Tool in Medical Education. Med Sci Educ. 2019;29(1):255-263. DOI: 10.1007/s40670-018-00687-w
21. Khan R, Payne MW, Chahine S. Peer assessment in the objective structured clinical examination: A scoping review. Med Teach. 2017;39(7):745-756. DOI: 10.1080/0142159X.2017.1309375
22. Basehore PM, Pomerantz SC, Gentile M. Reliability and benefits of medical student peers in rating complex clinical skills. Med Teach. 2014;36(5):409-414. DOI: 10.3109/0142159X.2014.889287
23. Hochlehnert A, Schultz JH, Möltner A, Timbil S, Brass K, Jünger J. Elektronische Erfassung von Prüfungsleistungen bei OSCE-Prüfungen mit Tablets. GMS Z Med Ausbild. 2015;32(4):Doc41. DOI: 10.3205/zma000983
24. Brennan RL. Generalizability Theory. New York NY: Springer; 2001. DOI: 10.1007/978-1-4757-3456-0
25. Epstein RM. Assessment in Medical Education. N Engl J Med. 2007;356(4):387-396. DOI: 10.1056/NEJMra054784

Korrespondenzadresse:
Andreas Möltner
Universität Heidelberg, Kompetenzzentrum für Prüfungen in der Medizin Baden-Württemberg, Im Neuenheimer Feld 346, 69120 Heidelberg, Deutschland
andreas.moeltner@med.uni-heidelberg.de

Bitte zitieren als
Möltner A, Lehmann M, Wachter C, Kurczyk S, Schwill S, Loukanova S. Formative assessment of practical skills with peer-assessors: quality features of an OSCE in general medicine at the Heidelberg Medical Faculty. GMS J Med Educ. 2020;37(4):Doc42. DOI: 10.3205/zma001335, URN: urn:nbn:de:0183-zma0013354

Artikel online frei zugänglich unter
https://www.egms.de/en/journals/zma/2020-37/zma001335.shtml

Eingereicht: *14.05.2019*
Überarbeitet: *24.03.2020*
Angenommen: *15.04.2020*
Veröffentlicht: *15.06.2020*