Fairness and objectivity of a multiple scenario objective structured clinical examination

Abstract

Introduction: The aim of the Objective Structured Clinical Examination (OSCE) is a standardized and fair assessment of clinical skills. Observing second clinical year medical students during a summative OSCE assessing a General Practice clerkship, we noticed that information exchange with peers led to a progressively faster and overly focused management of simulations. Therefore, we established a Multiple Scenario-OSCE (MS-OSCE) where all students had to manage the same chief complaint at a station but it's underlying scenarios being randomly changed during students' rotation through their parcours. We wanted to ensure they fully explore differential diagnosis instead of managing their task influenced by shared information. We wanted to assess if a MS-OSCE violates the assumption of objectivity and fairness given that students are not tested with the same scenarios.

Methods: We developed and piloted five OSCE stations (chest pain, abdominal pain, back pain, fatigue and acute cough) with two or three different underlying scenarios each. At each station these scenarios randomly changed from student to student. Performance was assessed with a checklist and global rating. The effect of scenarios and raters on students' grades was assessed calculating the intraclass correlation coefficient with a fixed effect two level linear model.

Results: A total of 169 students and 23 raters participated in the MS-OSCE. The internal consistency over all stations was 0.65 by Cronbach's alpha. The difference of the mean grades between the scenarios of a given chief complaint ranged from 0.03 to 0.4 on a 1 to 5 grading scale. The effect of scenarios on the variance of the final grades at each station ranged from 4% to 9% and of raters from 20% to 50% when adjusted for students' skills.

Conclusions: The effect of different scenarios on the grades was relevant but small compared to the effect of raters on grades. Improving rater training is more important to ensure objectivity and fairness of MS-OSCE than providing the same scenario to all students.

Keywords: medical students, medical education, objective structured clinical examination, rater effects

Johannes Spanke[1]
Christina Raus[1]
Annekathrin Haase[1]
Aniela Angelow[1]
Fabian Ludwig[1]
Gesine Weckmann[1,2]
Carsten Oliver Schmidt[3]
Jean-Francois Chenot[1]

1 University Medicine Greifswald, Institute for Community Medicine, Department of General Practice and Family Medicine, Greifswald, Germany

2 European University of Applied Sciences, Faculty of Applied Health Sciences, Rostock, Germany

3 University Medicine Greifswald, Institute for Community Medicine, SHIP-KEF, Greifswald, Germany

Introduction

The Objective Structured Clinical Examination (OSCE) is a common method to assess clinical and procedural skills in undergraduate medical education since its introduction by Harden et al. in 1975 [1]. We assess the clerkship in General Practice of second clinical year medical students with a summative OSCE. Standardized patients (SP) are used in OSCEs to ensure that each student encounters identically portrayed scenarios [2], [3]. As inherent to any assessment of clinical competence, objectivity (i.e. validity, reliability, efficiency, transparency) is susceptible to implementation and realisation imperfections [4], [5], [6], [7]. Additionally, cheating during OSCEs poses a threat to objectivity and fairness [8], [9], [10]. Fairness is the quality of making judgements that are free from bias and discrimination and requires conformity rules and standards for all students [11].

We assume that exchange of detailed information about the content of the OSCE-stations might be the cause for observations we made in previous years: It takes three days to assess the entire cohort of second clinical year medical students. We noticed that many students scheduled after the first round managed OSCE-simulations progressively faster and disproportionally focused. They often jumped to conclusions based on information they did not elicit during the simulation. For example, they made diagnosis and management decisions without having completed physical examination and history taking. As "communication skills" on electronic platforms are common among modern-day students, the sharing of information about the content of exams has become easier

[12], [13]. We identified internet blogs from medical students who finished the OSCE, providing hints to other students. We observed that students used case-specific information during ongoing examinations. Although several studies found that this kind of cheating does not necessarily effect test results to a relevant extent [9], [10], [14], [15], we believe this had a negative effect on the performance of students during examination.
Therefore, we established a Multiple Scenario-OSCE (MS-OSCE) where all students had to manage the same chief complaint with different underlying scenarios. The goal of multiple scenarios is to ensure that all students take a thorough history and perform a complete physical examination to explore the differential diagnoses at each OSCE-station, despite prior information received from students who already completed the OSCE. Varying an OSCE station while students are rotating on their examination parcours seems to be frequently done but has not been published extensively, whereas the effects of changing raters during an examination is well documented [16].
The aim of our analysis was to asses if a MS-OSCE violates the assumption of objectivity and fairness, given that all students are not tested with identical scenarios. Our hypothesis is that testing the management of a chief complaint with multiple scenarios does not unfairly affect the grading of students' performance.

Methods

This is an observational study about the implementation of the MS-OSCE concept to assess the General Practice clerkship of 169 second clinical year medical students (58% female, median age 26 years, range 22 to 37) [17]. Two students dropped out due to sickness.

Development of the MS-OSCE stations

In accordance with the competencies and learning objectives of the General Practice curriculum we generated an OSCE blueprint and developed five OSCE- stations, each testing one chief complaint with two to three different scenarios. Chief complaints for the OSCE were published on the website of the department of General Practice four weeks in advance to the OSCE to allow students to prepare for the examination. Chief complaints were: chest pain, abdominal pain, back pain, fatigue and acute cough. There are national guidelines for managing these complaints except for abdominal pain. The chief complaints with the respective underlying scenarios are summarized in table 1. The multiple scenarios chest pain station had been piloted in the previous year OSCE. The other OSCE-stations have been piloted with volunteer students.

Simulation patients and rater training

The scenarios for each chief complaint were standardized. Theatre students and lay-actors were recruited as simulation patients (SP). SPs were instructed to use a standardized opening phrase and received a detailed script describing the standardized way of interacting for each scenario (see table 1). We rehearsed the simulation with advanced medical students and physicians in postgraduate training. Elderly SPs simulated all chest pain scenarios for a more realistic portrayal of a possible cardiac origin of chest pain. The elderly chest pain SPs were trained portraying acute coronary syndrome previously and received additional training for costosternal syndrome and gastrointestinal reflux. Male SPs exclusively portrayed the abdominal pain scenarios to exclude gynaecological differential diagnoses. SPs completed a four hours training, including a rehearsal for every scenario with house officers.
Raters were General Practitioners (GPs) from the teaching practices network of the faculty. Most of them have been involved in rating OSCE for many years. All received a 15-30 minutes introduction to the new principles of the MS-OSCE before making their first assessment. The checklist for each chief complaint was identical. The scenarios were recapitulated with the SPs. Each station was assessed by 1 rater. During the three days of examination 23 raters were engaged. Two raters rated at all stations while most raters only rated at one or two stations.
Students enrolled electronically for a specific day and time slot. They were assigned to 2 groups of 5 students each. Two groups simultaneously circulated through a 5 stations course in a corridor with 10 separate rooms. The scenario to be simulated was randomly selected by the rater before the student entered the station. Students had 10 minutes at each station to complete the task and additional time to switch between stations. The entire MS-OSCE took 60 minutes for every student.

Assessment and grading

Federal regulations of examination in medical education in Germany require grading on an ordinal scale ranging from 1 to 5 (excellent (1), good (2), fair (3), sufficient (4) and fail (5)). This scale is used in a similar way in German schools and is familiar to all raters [https://www.gesetze-im-internet.de/_appro_2002/BJNR240500002.html]. We assessed students' performance with a checklist (checklist rating (CR)), which consisted of either binary items (e.g. student asked about smoking: yes/no) or Likert scales (e.g. quality of student–patient interaction). Checklist-items covered an identical examination routine for each scenario of a chief-complaint. Items fulfilled by more than 90% or less than 10% of the students were eliminated post hoc from the checklist. Communication was assessed with the Berlin Global Rating Scale grade (BGR) [18], a global rating scale [19], [20] based on the rating scale introduced by Hodges [21], adapted and validated for German assessment needs. Finally, raters had to give their intuitive overall global rating (OGR) [22] of each student's overall performance at each station. OGR is needed to calibrate CR and BGR for aspects that are not captured by the checklist. The

Table 1: Chief complaints with matching scenarios

Chief complaint	chest pain			lower back pain		abdominal pain			fatigue		cough		
SP sex	male/female			male		Male			female		male/female		
SP age	>50			18-30		18-30			18-30		18-50		
opening sentence	Doctor, I have chest pain.			When I got up after tying my shoes this morning, i felt a shooting poain in my back.		I don't fell well. Do you have a remedy for abdominal pain? I can't take it any longer.			Doctor, I'm tired all the time.		I have ab bad cough.		
Aetiology	Titze syndrome	gastric reflux	angina pectoris	uncom-plicated	herniated disc	appen-dicitis	chole-lithiasis	gastro-enteritis	hypo-thyroidism	depression	pneumonia	bronchitis	asthma
selected symptoms	pain on pressure on the costos-ternal border	burning retros-ternal pain, asso-ciated with food Ingestion	pain radiatring into the left arm	lumbar non-radiatring pain	radiation of pain tot he leg, weak-ness in the foot	guarding right lower quadrant pain	upper abdominal pain	upper abdominal pain	freezing consti-pation	difficulties sleeping, guilt, anhedonia	crackles weakness sweating	normal-lung sounds	wheez-ing dry cough dy-spnoea

SP simulation patient

final grade for each station was calculated as the mean of CR, BGR and OGR. According to the examination regulations at the University of Greifswald, a pre-fixed cut-off-score of 60% was set as standard for failure.

Statistical analysis

We display grades across scenarios as box-plots with average, median, interquartile range, and outliers. The internal consistency of the OSCE was assessed with Cronbach's alpha, based on the grades at each station. We computed intraclass correlations (ICC) to express the fraction of variance of the grade due to scenarios or raters. Ideally the fraction should be close to zero. For this purpose we computed linear regression models separately for each station, using a bootstrap approach for variance estimation because of violations of the normal distributions of the residuals. We used two sets of predictors:

1. dummy coded scenarios and raters (see table 2);
2. the first model and additionally the mean grade from all stations other than the outcome station (see table 3).

The grades were included to correct for students' overall skills on all stations except for the station under study. Computations were conducted with the xtreg command in stata, using the fixed-effects estimator. There were no missing data for the assessed variables.
Analyses were conducted in Stata 13 (Stata Corp., College Station, TX).

Results

Stations and raters as well as scenarios were statistically independent of each other (see attachment 1 and attachment 2). The internal consistency of the OSCE according to Cronbach's alpha across the five grades for the stations was 0.65 ($CI_{90\ one\ sided}$ 0.59).

Comparison of the scenarios at each OSCE Station

The distribution of grades for each scenario within stations and the distribution of final grades derived from the grades at each station are shown in figure 1. The average grade at each station ranged from 2.16 to 2.28. The difference of the average grade between the scenarios at each station ranged from 0.03 to 0.40 (see table 2 and table 3). The largest difference was observed at the station assessing chest pain management. The life-threatening scenario ACS had a worse average grade of 0.4 compared to the scenario of gastrointestinal reflux. A similar moderately worse grade of 0.3 was observed for the scenario of appendicitis compared to gastroenteritis. The final grades for the chief complaints (stations) ranged from 1 to 5.

Effect of scenarios and raters on the grades at each station

The effect of scenarios and raters on the grades at each station are expressed as ICCs and displayed in table 2 and table 3. We report the ICC unadjusted for students' skills (see table 2) and the ICC adjusted for students' skills at the other OSCE-stations (see table 3). The effect of the scenarios on the grades at the stations ranged from 5.2% to 7.8% without taking mean grades at the other stations into account and adjusted from 4.2% to 9.2% when taking the mean grade into account. Corresponding to the largest difference in average grades between the scenarios, the largest effect of scenario was observed at the station assessing chest pain.
The number of raters at each station varied from 6 to 10 over the three days. The unadjusted effect of the raters on the grades at the stations ranged from 14.1% to 39.8% without taking mean grades at the other stations into account and from 20.5% to 50.3% if doing so. The largest effect of raters was observed at the station assessing abdominal pain.

Discussion

Summary of the main results

A total of 169 second clinical year students and 23 raters participated in the MS-OSCE. The difference of the mean grades between the scenarios of a given chief complaint ranged from 0.03 to 0.4 on a 1-5 grading scale. The effect of scenarios on students' grades at a station accounted for 4% to 9% of the total variability of the grades, the re-

Table 2: Effect of station and raters on the grade unadjusted for students' skills

OSCE station	Scenario	Frequency n (%)	number of different raters	average grade of each scenario	overall average grade at station	most lenient rater average grade (n = rated students)	most strict rater average grade (n = rated students)	ICC for scenario	ICC for rater
abdominal pain	gastroenteritis	57 (33.73)	10	2.00	2.16	1.46 (15)	3.07 (20)	5.9%	39.8%
	appendicitis	56 (33.14)		2.33					
	bilious attack	56 (33.14)		2.14					
cough	bronchitis	57 (33.73)	7	2.27	2.20	1.75 (29)	2.63 (5)	6.2%	20.5 %
	pneumonia	58 (34.32)		2.30					
	asthma	54 (31.95)		2.01					
chest pain	acute coronary syndrome	59 (34.91)	6	2.52	2.28	1.98 (84)	2.85 (30)	7.8 %	14.1 %
	costosternal syndrome	56 (33.14)		2.19					
	reflux	54 (31.95)		2.12					
tiredness	psychogenic cause	80 (47.34)	8	2.37	2.24	1.56 (30)	2.88 (5)	5.2 %	25.7 %
	somatic cause	89 (52.66)		2.14					
low back pain	non-specific low back pain	89 (52.66)	7	2.02	2.16	1.51 (15)	2.79 (50)	5.8 %	30.2 %
	radicular pain	80 (47.34)		2.32					

ICC: intraclass correlation coefficient

Table 3: Effect of station and raters on the grade adjusted for students' skills

OSCE station	Scenario	Frequency n (%)	number of different raters	average grade of each scenario	overall average grade at station	most lenient rater average grade (n = rated students)	most strict rater average grade (n = rated students)	ICC for scenario	ICC for rater
abdominal pain	gastroenteritis	57 (33.73)	10	2.00	2.16	1.46 (15)	3.07 (20)	7.2 %	50.3 %
	appendicitis	56 (33.14)		2.33					
	bilious attack	56 (33.14)		2.14					
cough	bronchitis	57 (33.73)	7	2.27	2.20	1.75 (29)	2.63 (5)	4.2 %	24.6 %
	pneumonia	58 (34.32)		2.30					
	asthma	54 (31.95)		2.01					
chest pain	acute coronary syndrome	59 (34.91)	6	2.52	2.28	1.98 (84)	2.85 (30)	9.2 %	20.5 %
	costosternal syndrome	56 (33.14)		2.19					
	reflux	54 (31.95)		2.12					
tiredness	psychogenic cause	80 (47.34)	8	2.37	2.24	1.56 (30)	2.88 (5)	4.5 %	29.2 %
	somatic cause	89 (52.66)		2.14					
low back pain	non-specific low back pain	89 (52.66)	7	2.02	2.16	1.51 (15)	2.79 (50)	5.6 %	28.1 %
	radicular pain	80 (47.34)		2.32					

ICC: intraclass correlation coefficient

spective figures for raters ranged from 20% to 50% adjusted for students' skills.

Meaning of the findings

We observed differences in the distribution of the final grades between the scenarios ranging from 0.03 to 0.4 on the 5-point rating system (see figure 1). Although the checklist-items cover an identical examination routine for each scenario, rating should not be affected by the severity of the portrayed underlying diagnosis, since we expect students to explore all possibilities. It seems that missing the diagnosis or committing management errors for a potentially life-threatening scenario like ACS, appendicitis or pneumonia resulted in worse grades than similar mistakes with a corresponding benign scenario as costosternal syndrome, gastroenteritis, or bronchitis. There is no consensus what is considered a meaningful difference; we consider the observed difference as minor to moderate.

Compared to the magnitude of the effect of different raters on the grades at a station the effect of the different scenario was small but still relevant. The effect of the raters was independent of the scenarios and students' ability. The difference in the average grade awarded between the most lenient and strictest rater exceeded more than 1 grade on the 5-point rating scale, suggesting possibly poor inter-rater reliability. Therefore, calibrating raters seems to be far more important than adjusting for differences in the difficulty of scenarios. Wilkinson et al. [23] showed "that examiner factors contribute substantially more to the objectivity of an OSCE than do mark sheets or checklists". Inter-rater reliability in OSCEs is rarely reported and varies according to OSCE construction, rating instrument used (global rating/checklist rating) and assessment conditions (direct observation/ video) [20], [24], [25]. Hatala et al. [26] piloted an OSCE with 2 stations, fragmented into 3 subsequent sequences of 10 minutes each to cover multiple content areas relevant to internal medicine. They observed an inter-rater reliability ranging from 0.63 to 0.91 with two raters for each scenario. Brennan and colleagues [16] found that although the range of grades awarded varied if examiners changed at OSCE stations (total number of raters at a given station not stated), examination reliability and the likely candidate outcome were not affected.

Due to financial constraints, we - like many other medical schools - cannot afford to assess each OSCE station with two raters simultaneously.

More intensive training of raters and SPs [4] as well as a more thorough development of checklists to establish better inter-rater reliability are possible remedies to reduce the effect of raters on grading. However, the assumption that a more intensive rater training increases inter-rater reliability does not always hold true [27], [28]. Which amount of unfairness and lack of reliability should be

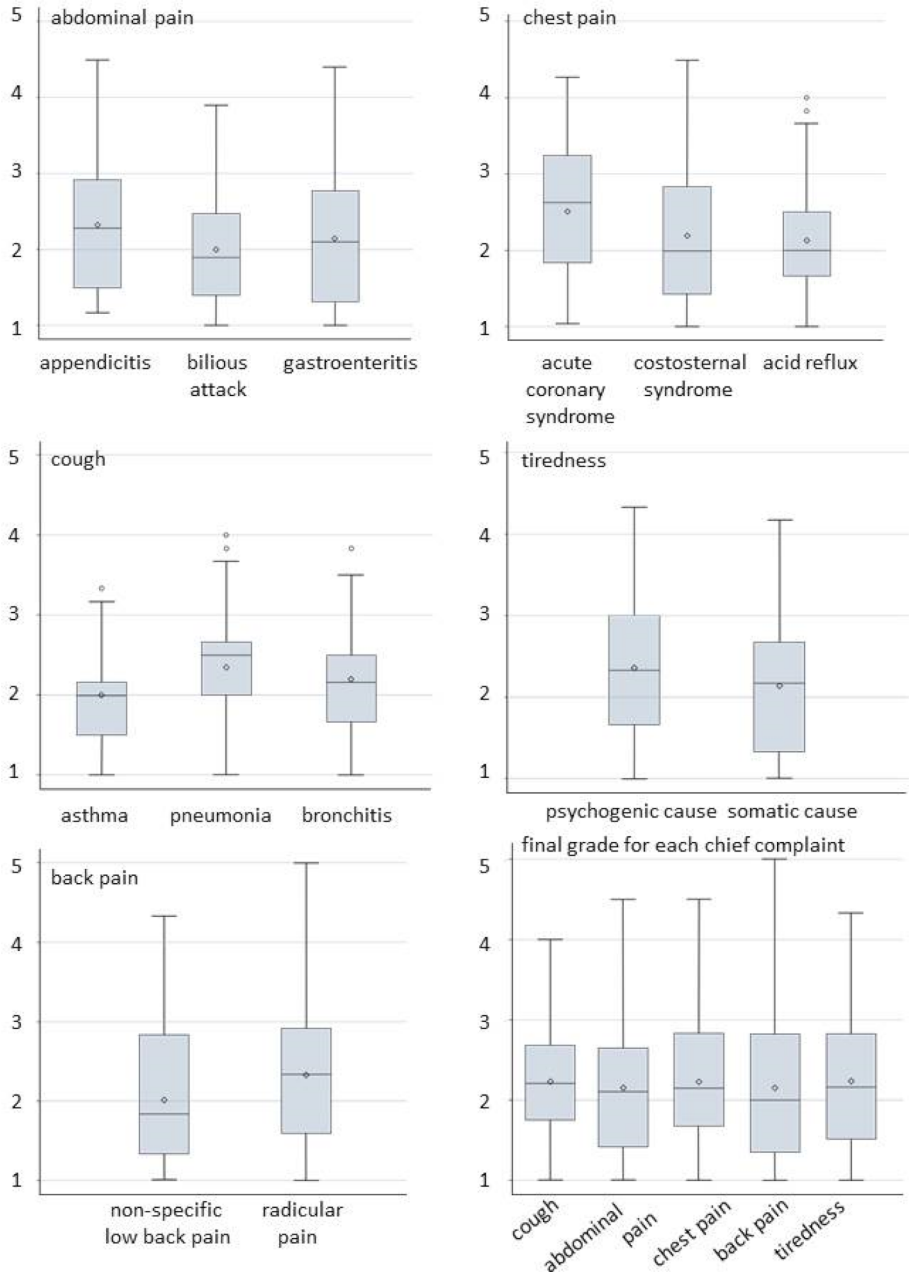

Figure 1: The distribution of the grades is displayed as box plots showing the median (horizontal line), average (diamond) and the interquartile range (lengths of the box). The vertical lines (whiskers) show minimum and maximum values excluding outliers, which are displayed as dots.

accepted and to which degree the effect of raters can be reduced is a matter of debate [29].

We do not believe that MS-OSCE has reduced exchange of information, but we assume subjectively that the switch to MS-OSCE has led to a more complete history taking and physical examination and a less hasty performance throughout the whole 3 days of the annual OSCE. However, we have no objective measurement supporting this assumption.

Strengths and limitations

This is to our knowledge the first report of a MS-OSCE. We calculated the impact of multiple scenarios and raters on the grades in a MS-OSCE adjusting for students' skills. We did not establish inter-rater correlations for the checklists and provided only minimal rater training, due to lack of resources. This reflects most likely the situation at many medical schools assessing students' skills with OSCE. There was a good correlation ranging from 0.6 to 0.8 between the checklist rating and global rating (results not shown), indicating congruent ratings of communication and examination skills. We cannot exclude effects on students' performance due to different accuracy in portrayal of scenarios by different SPs portraying the same scenario during three days of examination. We did not attempt to adjust for SPs. Additionally we did not investigate or adjust for gender effects which have been shown to effect grading [29], [30], [31]. Varying gender of SPs might have influenced students' performance at

the chest pain station and the acute cough station, where auscultation was within the scope of the demanded skills. Our MS-OSCE with only five stations is relatively short. It has been postulated that at least 10 stations are needed for a reliable assessment [32], [33]. Ten minutes per station is in an accepted time range [34], [35] and even high-stakes examinations demand only 15 minutes per OSCE-station for patient encounter [36]. We have a good internal consistency (Cronbach's alpha: 0.65) over all stations compared with other reports from the literature [32].

Although it is possible to adjust students' individual grades for differences in scenario and for differences between raters with a correction factor after taking the exam, we did not adjust accordingly.Calculation of correction factors after each exam would require resources which are currently not available to us.

Validity measurements are not in the scope of our report. Van der Vleuten and Schuwirth [7] state that key issues concerning the validity of competence assessments are authenticity of performance and the integration of professional competencies. MS-OSCE addresses the authenticity of students' performance by providing several scenarios at one station to reduce the effect of shared information (cheating) on students' case management. Content validity was assured by reviewing MS-OSCE-stations by a team of experienced teaching physicians. Providing SP-based clinical scenarios at each station, assessment by standardised ratings (checklists) and a validated global rating instrument, face validity of the MS-OSCE might equal that of a traditional OSCE with only 5 stations.

Conclusions

The effect of different scenarios on the grades assessing the management of one chief complaint in General Practice was small compared to the effect of raters. Improving inter-rater reliability is more important to ensure objectivity and fairness of OSCE than providing the same scenario to all students.

List of abbreviations

ACS: acute coronary syndrome
BGR: Berlin Global Rating Scale
CI: confidence interval
CR: checklist rating
GP: General Practitioner
OGR: overall global rating
ICC: intraclass correlation coefficient
MS-OSCE: Multiple Scenario Objective Structured Clinical Examination
OSCE: Objective Structured Clinical Examination
SP: Standardized Patient

Acknowledgements

We are grateful to Francis Baudet, Gisela Greschniok, Heinz Hammermayer, Thomas Hannemann, Mathias Herberg, Gero Kärst, Andreas Krüger, Barbara Krüger, Annika Matz, Hans-Diether Seiboth, Thomas Richter, Claudia Runge, Carmina Spreemann, Antje Theurer, Renate Tilchner, Rüdiger Titze, Arne Wasmuth, Christine Wendt, Arno Wilfert.

Data availability

Data is available on reasonable request.

Authors' contributions

JS and JFC conceived the multiple scenario OSCE, the scenarios and rating sheets were developed and piloted by JS, CR, GW, AA, FL, AH, JFC. CR, JS, FL and GW trained the simulation patients, AH was responsible for data management, COS was leading the statistical analysis. JS and JFC wrote the first draft which was revised and approved by all authors.

Competing interests

The authors declare that they have no competing interests.

Attachments

Available from
http://www.egms.de/en/journals/zma/2019-36/zma001234.shtml

1. Attachment_1.pdf (219 KB)
   Independence of stations and raters. Information attachment 1: to assess independence Chi² or Fisher exact test was calculated.
2. Attachment_2.pdf (233 KB)
   Independence of scenarios. Information attachment 2: to assess independence Chi² or Fisher exact test was calculated.

References

1. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J. 1975;1(5955):447-451. DOI: 10.1136/bmj.1.5955.447
2. Vu NV, Barrows HS. Use of Standardized Patients in Clinical Assessments: Recent Developments and Measurement Findings. Educ Res. 1994;23:23-30. DOI: 10.3102/0013189X023003023
3. Patrício MF, Julião M, Fareleira F, Carneiro AV. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? Med Teach. 2013;35(6):503-514. DOI: 10.3109/0142159X.2013.774330

4. Baig LA, Beran TN, Vallevand A, Baig ZA, Monroy-Cuadros M. Accuracy of portrayal by standardized patients: results from four OSCE stations conducted for high stakes examinations. BMC Med Educ. 2014;14:97. DOI: 10.1186/1472-6920-14-97

5. Van der Vleuten CP, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. Med Educ. 1991;25(2):110-118. DOI: 10.1111/j.1365-2923.1991.tb00036.x

6. Furman GE, Smee S, Wilson C. Quality assurance best practices for simulation-based examinations. Simul Healthc. 2010;5(4):226-231. DOI: 10.1097/SIH.0b013e3181da5c93

7. van der Vleuten CP, Schuwirth LW. Assessing professional competence. From methods to programmes. Med Educ. 2005;39(3):309-317. DOI: 10.1111/j.1365-2929.2005.02094.x

8. Parks R, Warren PM, Boyd KM, Cameron H, Cumming A, Lloyd-Jones G. The Objective Structured Clinical Examination and student collusion: marks do not tell the whole truth. J Med Ethics. 2006;32(12):734-738. DOI: 10.1136/jme.2005.015446

9. Colliver JA, Barrows HS, Vu NV, Verhulst SJ, Mast TA, Travis TA. Test security in examinations that use standardized-patient cases at one medical school. Acad Med. 1991;66(5):279-282. DOI: 10.1097/00001888-199105000-00011

10. Colliver JA, Travis TA, Robbs RS, Barnhart AJ, Shirar LE, Vu NV. Test security in standardized-patient examinations: analysis with scores on working diagnosis and final diagnosis. Acad Med. 1992;67(10):S7-S9. DOI: 10.1097/00001888-199210000-00022

11. Harden RM, Lilley P, Patricio M. The definitive guide to the OSCE: The Objective Structured Clinical Examination as a performance assessment. Edinburgh, New York: Elsevier; 2016.

12. Kennedy G, Gray K, Tse J. 'Net Generation' medical students: technological experiences of pre-clinical and clinical students. Med Teach. 2008;30(1):10-16. DOI: 10.1080/01421590701798737

13. Pander T, Pinilla S, Dimitriadis K, Fischer MR. The use of Facebook in medical education - a literature review. GMS Z Med Ausbild. 2014;31(3):Doc33. DOI: 10.3205/zma000925

14. Rutala PJ. Sharing of Information by Students in an Objective Structured Clinical Examination. Arch Intern Med. 1991;151(3):541. DOI: 10.1001/archinte.1991.00400030089016

15. Wilkinson TJ, Fontaine S, Egan T. Was a breach of examination security unfair in an objective structured clinical examination? A critical incident. Med Teach. 2003;25(1):42-46. DOI: 10.1080/0142159021000061413

16. Brennan PA, Croke DT, Reed M, Smith L, Munro E, Foulkes J, Arnett R. Does Changing Examiner Stations During UK Postgraduate Surgery Objective Structured Clinical Examinations Influence Examination Reliability and Candidates' Scores? J Surg Educ. 2016;73(4):616-623. DOI: 10.1016/j.jsurg.2016.01.010

17. Chenot JF. Undergraduate medical education in Germany. GMS Ger Med Sic. 2009;7:Doc02. DOI: 10.3205/000061

18. Scheffer S. Validierung des "Berliner Global Rating" (BGR). Ein Instrument zur Prüfung kommunikativer Kompetenzen Medizinstudierender im Rahmen klinisch-praktischer Prüfungen (OSCE) [An instrument for assessing communicative competencies of medical students within the frame of testing clinical skills]. Berlin: Charité - Universitätsmedizin Berlin, Medizinische Fakultät; 2009. Zugänglich unter/available from: http://nbn-resolving.de/urn:nbn:de:kobv:188-fudissthesis000000010951-7

19. Regehr G, Freeman R, Robb A, Missiha N, Heisey R. OSCE performance evaluations made by standardized patients: comparing checklist and global rating scores. Acad Med. 1999;74(10 Suppl):S135-S137. DOI: 10.1097/00001888-199910000-00064

20. Ilgen JS, Ma IWY, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. Med Educ. 2015;49(2):161-173. DOI: 10.1111/medu.12621

21. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Med Educ. 2003;37(11):1012-1016. DOI: 10.1046/j.1365-2923.2003.01674.x

22. Hunter DM, Jones RM, Randhawa BS. The use of holistic versus analytic scoring for large-scale assessment of writing. Can J Prog Eval. 1996;11:61-85.

23. Wilkinson TJ, Frampton CM, Thompson-Fawcett M, Egan T. Objectivity in Objective Structured Clinical Examinations: Checklists Are No Substitute for Examiner Commitment. Acad Med. 2003;78(2):219-223. DOI: 10.1097/00001888-200302000-00021

24. Kim J, Neilipovitz D, Cardinal P, Chiu M. A comparison of global rating scale and checklist scores in the validation of an evaluation tool to assess performance in the resuscitation of critically ill patients during simulated emergencies (abbreviated as "CRM simulator study IB"). Simul Healthc. 2009;4(1):6-16. DOI: 10.1097/SIH.0b013e3181880472

25. Malau-Aduli BS, Mulcahy S, Warnecke E, Otahal P, Teague PA, Turner R, van der Vleuten C. Inter-Rater Reliability: Comparison of Checklist and Global Scoring for OSCEs. Creat Educ. 2012; 03:937-942. DOI: 10.4236/ce.2012.326142

26. Hatala R, Marr S, Cuncic C, Bacchus CM. Modification of an OSCE format to enhance patient continuity in a high-stakes assessment of clinical performance. BMC Med Educ. 2011;11:23. DOI: 10.1186/1472-6920-11-23

27. Weitz G, Vinzentius C, Twesten C, Lehnert H, Bonnemeier H, König IR. Effects of a rater training on rating accuracy in a physical examination skills assessment. GMS Z Med Ausbild. 2014;31(4):Doc41. DOI: 10.3205/zma000933

28. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores. A randomized, controlled trial. J Gen Int Med. 2009;24(1):74-79. DOI: 10.1007/s11606-008-0842-3

29. Schleicher I, Leitner K, Juenger J, Moeltner A, Ruesseler M, Bender B, Sterz J, Schuettler KF, Koenig S, Kreuder JG. Examiner effect on the objective structured clinical exam - a study at five medical schools. BMC Med Educ. 2017;17(1):71. DOI: 10.1186/s12909-017-0908-1

30. Mortsiefer A, Karger A, Rotthoff T, Raski B, Pentzek M. Examiner characteristics and interrater reliability in a communication OSCE. Pat Educ Couns. 2017;100(6):1230-1234. DOI: 10.1016/j.pec.2017.01.013

31. Carson JA, Peets A, Grant V, McLaughlin K. The effect of gender interactions on students' physical examination ratings in objective structured clinical examination stations. Acad Med. 2010;85(11):1772-1776. DOI: 10.1097/ACM.0b013e3181f52ef8

32. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Med Educ. 2011;45(12):1181-1189. DOI: 10.1111/j.1365-2923.2011.04075.x

33. Nikendei C, Jünger J. OSCE - hands on instructions for the implementation of an objective structured clinical examination. GMS Z Med Ausbild. 2006;23(3):Doc47. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2006-23/zma000266.shtml

34. Reznick RK, Blackmore D, Dauphinee WD, Rothman AI, Smee S. Large-scale High-stakes Testing with an OSCE: Report from the Medical Council of Canada. Acad Med. 1996;71(1 Suppl):S19-S21. DOI: 10.1097/00001888-199601000-00031

35. Hamann C, Volkan K, Fishman MB, Silvestri RC, Simon SR, Fletcher SW. How well do second-year students learn physical diagnosis? Observational study of an objective structured clinical examination (OSCE). BMC Med Educ. 2002;2:1-11. DOI: 10.1186/1472-6920-2-1

36. Chambers KA, Boulet JR, Gary NE. The management of patient encounter time in a high-stakes assessment using standardized patients. Med Educ. 2000;34(10):813-817. DOI: 10.1046/j.1365-2923.2000.00752.x

Corresponding authors:
Johannes Spanke
University Medicine Greifswald, Institute for Community Medicine, Department of General Practice and Family Medicine, Fleischmannstr. 6, D-17475 Greifswald, Germany
johannes.spanke@uni-greifswald.de
Christina Raus
University Medicine Greifswald, Institute for Community Medicine, Department of General Practice and Family Medicine, Fleischmannstr. 6, D-17475 Greifswald, Germany

Please cite as
Spanke J, Raus C, Haase A, Angelow A, Ludwig F, Weckmann G, Schmidt CO, Chenot JF. Fairness and objectivity of a multiple scenario objective structured clinical examination. GMS J Med Educ. 2019;36(3):Doc26.
DOI: 10.3205/zma001234, URN: urn:nbn:de:0183-zma0012343

This article is freely available from
http://www.egms.de/en/journals/zma/2019-36/zma001234.shtml

Received: *2018-05-23*
Revised: *2018-11-11*
Accepted: *2019-02-13*
Published: *2019-05-16*

OPEN ACCESS

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

Gerechtigkeit und Objektivität einer OSCE-Prüfung mit multiplen Szenarien

Johannes Spanke[1]
Christina Raus[1]
Annekathrin Haase[1]
Aniela Angelow[1]
Fabian Ludwig[1]
Gesine Weckmann[1,2]
Carsten Oliver Schmidt[3]
Jean-Francois Chenot[1]

1 University Medicine Greifswald, Institute for Community Medicine, Department of General Practice and Family Medicine, Greifswald, Deutschland

2 Europäische Fachhochschule Rhein/Erft, Fachbereich Angewandte Gesundheitswissenschaften, Rostock

3 University Medicine Greifswald, Institute for Community Medicine, SHIP-KEF, Greifswald, Deutschland

Zusammenfassung

Hintergrund: Das Ziel einer Objective Structured Clinical Examination (OSCE-Prüfung) ist eine standardisierte und faire Prüfung klinischer Fertigkeiten. Nach dem Blockpraktikum Allgemeinmedizin im 2. klinischen Jahr (4. Studienjahr) werden die Studierenden mit einer OSCE-Prüfung Allgemeinmedizin an Simulationspatienten beurteilt. In der Vergangenheit konnten wir beobachten, dass prüfungsrelevante Informationen während der Prüfung unter den Studierenden ausgetauscht wurden. Dies führte zu einer zunehmend hastigen und unpräzisen Interaktion mit dem Simulationspatienten. Daher entwickelten wir eine Multiple-Scenario-OSCE-Prüfung (MS-OSCE), bei der an jeder Station einem bestimmten Beratungsanlass unterschiedliche Szenarios zugrunde gelegt werden, die bei gleichlautender Aufgabenstellung während der Rotation einer Studierendengruppe innerhalb jeder Station randomisiert gewechselt wurden. Eine MS-OSCE soll die Studierenden veranlassen, mögliche Differentialdiagnosen gründlicher zu explorieren, anstatt ihre Aufgaben unter dem Einfluss von weitergeleiteten Informationen vorangehender Prüfungskandidaten zu lösen. Wir wollten beurteilen, ob die unterschiedlichen Szenarien einer Station vergleichbare Schwierigkeiten aufwiesen und welche Faktoren die Fairness und Objektivität der MS-OSCE beeinflussen.
Methoden: Wir entwickelten und pilotierten fünf OSCE-Stationen (Beratungsanlässe: Brustschmerz, Bauchschmerz, Rückenschmerz, Müdigkeit und akuter Husten) mit zwei oder drei unterschiedlichen Szenarien für den an der jeweiligen Station vorgesehenen Beratungsanlass. Der Wechsel der Szenarios an jeder Station erfolgte randomisiert von Student/in zu Student/in. Die Leistungsbewertung der Studierenden erfolgte sowohl mit einer Checkliste als auch mit einem globalen Rating. Der Effekt der Szenarien und der Prüfer/-in auf die Noten der Studierenden wurde durch Berechnung des Intraclass-Korrelationskoeffizienten mit einem linearen Zweiebenen-Modell mit fixen Effekten ermittelt.
Ergebnisse: An der MS-OSCE nahmen insgesamt 169 Studierende und 23 Prüfer/innen teil. Die mittels Cronbach's alpha berechnete Interne Konsistenz über alle Stationen auf einer Notenskala von 1 bis 5 betrug 0,65. Die mittlere Notendifferenz zwischen den Szenarien eines Beratungsanlasses reichte von 0,03 bis 0,4. Der Einfluss der Szenarien auf die Varianz der durchschnittlichen Noten pro Station lag nach Adjustierung für die Fähigkeiten der Studierenden bei 4% bis 9%. Der Einfluss der Prüfer/-innen reichte von 20% bis 50%.
Schlussfolgerung: Der Einfluss der unterschiedlichen Szenarien einer Station auf die Note war gering im Vergleich zum Einfluss der Prüfer/-in. Um die Objektivität einer MS-OSCE zu gewährleisten muss eine adäquate Prüferschulung erfolgen. Verbesserung der Interrater-Reliabilität ist wichtiger für Fairness und Objektivität, als alle Studierenden mit demselben Szenario zu prüfen.

Schlüsselwörter: Medizinstudenten, Medizinische Ausbildung, OSCE, Prüfereffekte

Einführung

Seit der Einführung durch Harden 1975 [1] hat sich die Objective Structured Clinical Examination (OSCE) zur Prüfung von klinischen Fähigkeiten und Fertigkeiten in der Ausbildung von Medizinstudierenden etabliert. Wir prüfen Medizinstudentinnen und Medizinstudenten nach ihrem Blockpraktikum Allgemeinmedizin im 2. Klinischen Jahr mit einer summativen OSCE-Prüfung. Mit geschulten Simulationspatienten (SP) wird bei einer OSCE-Prüfung, jede/r Studierende in standardisiert dargestellten klinischen Situationen geprüft [2], [3]. Die Objektivität von Prüfungen klinischer Kompetenz (Kriterien für Objektivität: Validität, Reliabilität, Effizienz, Transparenz) ist allerdings häufig beeinträchtigt durch Schwächen bei Planung und Durchführung der Prüfungen [4], [5], [6], [7]. Auch das Weitergeben von prüfungsrelevanten Informationen durch Studierende während einer OSCE-Prüfung stellt eine Beeinträchtigung von deren Fairness und Objektivität dar [8], [9], [10]. Fairness einer Prüfung bedeutet, dass Beurteilungen frei von Voreingenommenheit erfolgen und niemanden benachteiligen. Sie erfordert die Einhaltung übereinstimmender Regeln und Standards für alle Studierenden [11].

Wir nehmen an, dass der Austausch detaillierter Informationen zu Inhalten der OSCE-Stationen während der Prüfung ein Grund für Beobachtungen ist, die wir in den letzten Jahren gemacht hatten: Zur Prüfung der gesamten Jahrgangskohorte der Studierenden im 2. Klinischen Jahr benötigen wir 3 Tage. In diesem Zeitraum konnten wir beobachten, dass Studierende, die die Prüfung erst nach der ersten Prüfungsgruppe antraten, die Aufgaben an den OSCE-Stationen zunehmend hastiger und weniger nachvollziehbar absolvierten. Sie zogen Schlüsse, die nicht auf Informationen beruhten, die sie während der Interaktion mit der Simulationspatientin/ dem Simulationspatienten herausgearbeitet hatten. Sie kamen zum Beispiel zu einer Diagnose oder zu einer Therapieentscheidung, ohne eine ausreichende körperliche Untersuchung oder die Anamnese abgeschlossen zu haben. Da heutzutage jeder Student über "Kommunikative Fähigkeiten" mittels elektronischer Medien verfügt, ist es einfacher, Informationen zum Prüfungsinhalt zeitnah auszutauschen [12], [13]. Wir identifizierten Internetblogs von Medizinstudierenden, die ihre OSCE-Prüfung bereits absolviert hatten, in denen Hinweise für nachfolgende Prüflinge enthalten waren. Auch bemerkten wir, dass Studierende fallspezifische Informationen während der laufenden Prüfung benutzten. Auch wenn mehrere Studien zeigen konnten, dass diese Art von Fehlverhalten die Prüfungsergebnisse nicht notwendigerweise relevant beeinflusst [9], [10], [14], [15], nehmen wir an, dass diese Informationen einen negativen Effekt auf die Prüfungsleistung der Studierenden haben.

Darum entwickelten wir eine Multiple Scenario-OSCE-Prüfung (MS-OSCE), bei der alle Studierenden den immer gleichbleibenden Beratungsanlass einer Station managen müssen, jedoch mit wechselnden zugrundeliegenden Szenarien (d.h. Ursachen). Multiple Szenarien bei gleichem Beratungsanlass sollen dafür sorgen, dass alle Studierenden eine gründliche Anamnese und eine adäquate Untersuchung ausführen, trotz Informationen zur Prüfung von Studierenden, die die MS-OSCE-Prüfung vorher absolviert hatten. Das Verändern einer OSCE-Station im Verlauf einer Prüfung scheint nicht unüblich zu sein, doch wurde darüber bisher nur wenig publiziert; wohingegen der Effekt von wechselnden Prüfer/innen während einer OSCE-Prüfung gut dokumentiert ist [16].

Das Ziel unserer Untersuchung war es, zu prüfen, ob Objektivität und Fairness der MS-OSCE-Prüfung angenommen werden dürfen, auch wenn nicht alle Studierenden mit dem identischen Szenario eines Beratungsanlasses geprüft wurden.

Unsere Hypothese lautet, dass die Notengebung für die Prüfungsleistung der Studierenden nicht unfair beeinflusst wird, wenn das Management eines Beratungsanlasses mittels multipler Szenarien getestet wird.

Methoden

Dies ist eine Beobachtungsstudie zur Implementierung des MS-OSCE Konzepts. Die Prüfung war Teil der Benotung des Blockpraktikums Allgemeinmedizin von 169 Studierenden im 2. Klinischen Jahr. (58% weiblich, Median Alter: 26 Jahre (22-37 Jahre) [17]. Zwei Studierende des Jahrgangs traten die Prüfung wegen Krankheit nicht an.

Entwicklung der MS-OSCE Stationen

In Übereinstimmung mit dem Lernzielkatalog des Faches Allgemeinmedizin erstellten wir einen OSCE-Blueprint und entwickelten daraus fünf OSCE-Stationen, von denen jede einen anderen Beratungsanlass abprüfte mit jeweils 2-3 zugrundeliegenden Szenarien pro Beratungsanlass. Die Beratungsanlässe wurden 4 Wochen vor der OSCE-Prüfung auf der Website der Abteilung Allgemeinmedizin bekanntgegeben, um den Studierenden eine Vorbereitung auf die Prüfung zu ermöglichen. Die Beratungsanlässe lauteten: „Brustschmerz", „Bauchschmerz", „Rückenschmerz", „Müdigkeit" und „akuter Husten". Für das Management der genannten Beratungsanlässe, außer für „Bauchschmerz", existieren nationale Leitlinien. Die Synopse der Beratungsanlässe mit den entsprechend zugeordneten Szenarien ist in Tabelle 1 dargestellt. Die Pilotierung der Brustschmerzstation erfolgte bereits für die OSCE-Prüfung des Vorjahres. Die übrigen OSCE-Stationen wurden mit Hilfe freiwilliger Studierender pilotiert.

Simulationspatienten und Prüfertraining

Die Szenarien für jeden Beratungsanlass wurden standardisiert erstellt. Als Simulationspatienten/Simulationspatientinnen (SP) wurden Studierende einer Theaterakademie und Laienschauspieler/innen rekrutiert. Die SPs wurden instruiert, ihre Simulation immer mit einem festgelegten Eingangssatz zu beginnen und erhielten ein

Tabelle 1: Beratungsanlässe mit dazugehörigen Szenarien

Beratungsanlass	Brustschmerz			Rückenschmerz		Bauchschmerz			Müdigkeit		Husten		
SP Geschlecht	männlich/weiblich			männlich		männlich			weiblich		männlich/weiblich		
SP Alter	>50			18-30		18-30			18-30		18-50		
Eröffnungssatz des SP	„Herr/ Frau Doktor, ich habe so Schmerzen in der Brust“			„Als ich heute Morgen meine Schuhe schnüren wollte, ist mir beim Aufstehen ein Schmerz in den Rücken gefahren“		„Ich fühl' mich gar nicht gut. Können Sie mir was gegen meine Bauchschmerzen aufschreiben? Ich halte das nicht länger aus!“			„Herr/Frau Doktor ich bin die ganze Zeit so müde!“		„Ich habe so einen schrecklichen Husten!“		
Ätiologie	Costosternales Syndrom (Tietze-Syndrom))	Reflux gastro-öso-phageal	Angina pect-oris ACS	Unkompli-zierte Lum-balgie	Band-schei-ben-prolaps	Append-izitis	Cholezysto lithiasis	Gastro-enteritis	Hypo-thyreose	Depression	Pneumonie	Bron-chitis	Asthma
ausgesuchte Symptome	Schmerz am costoster-nalen Übergang	Bren-nender retro-sternaler Schmerz in Verbin-dung mit Essen	Schmerz mit Ausstrah-lung in den linken Arm	Lumbaler Schmerz ohne Ausstrah-lung	Schmerz Ausstrah-lung in das Bein, Schwäche im Fuß	Schmerz mit Abwehr-spannung im rechten Unter-bauch	Ober-bauch-schmerzen	Ober-bauch-schmerzen	Frösteln, Ver-stopfung	Schlaf-probleme, Schuld-gefühl, Lust-losigkeit	Rassel-geräusche, Schwäche, Schwitzen	Normale Lungen-aus-kultation	Giemen, trockener Husten, Luftnot

SP=Simulationspatient, ACS=akutes Koronarsyndrom

detailliertes Skript, in dem jedes Szenario eines Beratungsanlasses mit standardisierten Regieanweisungen beschrieben wurde (siehe Tabelle 1). Die einzelnen Szenarien wurden dann mit Medizinstudierenden ab dem dritten klinischen Jahr und Ärzten in Weiterbildung eingeübt. Für die Szenarien der „Brustschmerz“-Station wurden ältere SPs eingesetzt, um eine möglichst realistische Darstellung des in Frage kommenden akuten Koronarsyndroms zu gewährleisten. Diese älteren SPs hatten bereits im vorangehenden Jahr das Training für die Darstellung eines akuten Koronarsyndroms erhalten und wurden nun zusätzlich für die weiteren Szenarien der „Brustschmerz“-Station (costosternales Syndrom und gastroösophagealer Reflux) trainiert. An der „Bauchschmerz“-Station wurden ausschließlich männliche SPs eingesetzt, um gynäkologische Differentialdiagnosen auszublenden. Jede/r SP erhielt ein 4-stündiges Training inclusive einer Probe mit Supervision durch eine/n Lehrbeauftragten.

Die Prüfer/innen waren Allgemeinärztinnen und Allgemeinärzte des Lehrärztenetzwerks der Abteilung für Allgemeinmedizin. Die meisten von ihnen sind bereits seit Jahren als Prüfer/innen in OSCE-Prüfungen tätig gewesen. Alle Prüfer/innen erhielten eine 15-30 minütige Einführung in die neuen Prinzipien der MS-OSCE-Prüfung bevor sie die erste Bewertung abgaben. Die Checkliste eines jeden Beratungsanlasses war für die ihm zugrundeliegenden Szenarien identisch. Die Szenarien wurden vor Beginn der Prüfung mit den SP nochmals durchgesprochen. Jede Station war mit einem Prüfer/in besetzt. Im Laufe der 3 Prüfungstage wurden 23 Prüfer/innen tätig. 2 Prüfer/innen wechselten durch alle Stationen, wohingegen die meisten Prüfer/innen nur an 1 oder 2 Stationen eingesetzt waren.

Die Studierenden konnten sich elektronisch für einen Prüfungstag und den Prüfzeitraum einschreiben. Sie wurden einer der beiden Prüfungsgruppen zu je 5 Stationen zugeteilt. Beide Prüfungsgruppen zirkulierten simultan von Station 1 bis Station 5 in einem Flur mit 10 getrennten Räumen (2 x Stationen 1-5). Bevor eine Studentin/ein Student den Raum einer Station betrat, wählte die/der Prüfer/in das zu simulierende Szenario nach Zufall aus. Die Studierenden hatten 10 Minuten Zeit, die Aufgabenstellung einer Station zu lösen und wechselten danach alle gleichzeitig die Station in einer festgelegten Zeit. Jede/r Studierende benötigte 60 Minuten für die komplette MS-OSCE Prüfung.

Beurteilung und Notenvergabe

Die Approbationsordnung in Deutschland verlangt eine Notenvergabe auf einer Ordinalskala von 1 bis 5 (sehr gut (1), gut (2), befriedigend (3), ausreichend (4) und mangelhaft (5)). Dieses Benotungssystem wird in deutschen Schulen in ähnlicher Weise benutzt und ist allen Prüfern vertraut [https://www.gesetze-im-internet.de/_appro_2002/BJNR240500002.html]. Wir prüften die Fertigkeiten der Studentinnen und Studenten mit einer Checkliste (checklist rating (CR)), die sowohl binäre Items (z.B. Studierende/r fragt nach Nikotinkonsum: ja/nein) als auch Likert Skalen (z.B. Qualität einer ausgeführten Student/in – Patient/in – Interaktion) beinhaltete. Die Items der Checkliste einer Station erfassten für jedes Szenario eines Beratungsanlasses die gleichen Untersuchungserfordernisse. Die kommunikative Kompetenz wurde mit dem Berliner Global Rating Instrument (BGR) [18] bewertet, einer Globalbewertungsskala [19], [20], basierend auf der von Hodges eingeführten Beurteilungsskala [21], die für deutsche Prüfungserfordernisse angepasst und validiert wurde. Abschließend wurde eine intuitive globale Gesamtbeurteilung (overall global rating (OGR)) [22] für die Gesamtleistung der/des Studierenden an seiner Station abgegeben. Diese wird benötigt, um CR und BGR auf Aspekte hin abzugleichen, die von Checklisten nicht erfasst werden. Die Gesamtnote an jeder Station wurde errechnet als der Durchschnitt aus CR, BGR und OGR. Items, die von mehr als 90% oder weniger als 10% der Studierenden erfüllt wurden, wurden post hoc von der Checkliste gestrichen. Entsprechend der Studienordnung der Universität Greifswald wurde eine Bestehensgrenze von 60% der maximalen erreichbaren Punktzahl im Voraus festgelegt.

Statistische Auswertung

Wir zeigen die Noten einer Station über alle Szenarien als Box-Plots mit Mittelwert, Median, Interquartilenabstand und Ausreißern (siehe Abbildung 1). Die interne Konsistenz der OSCE-Prüfung wurde mittels Cronbach's

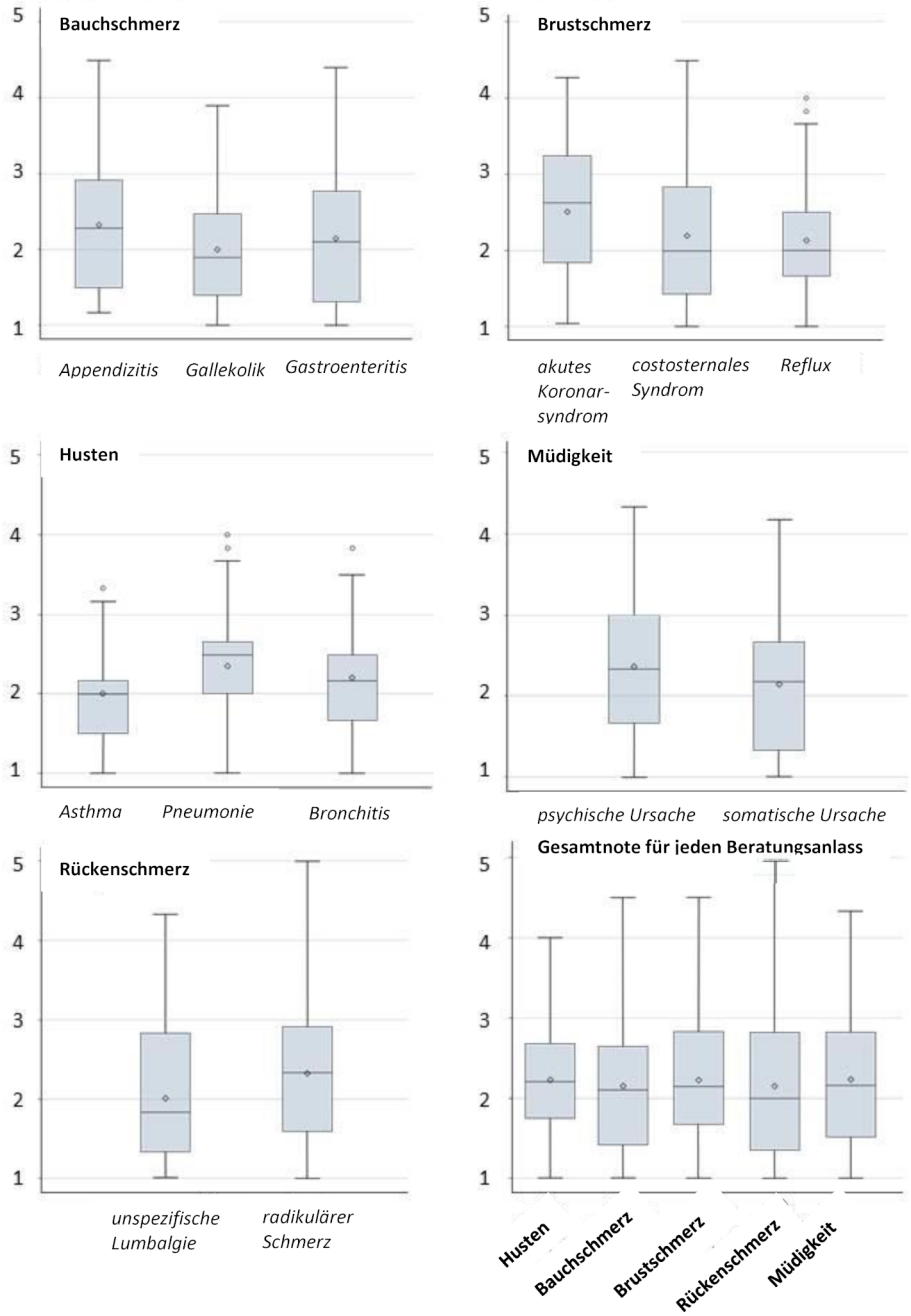

Abbildung 1: Die Verteilung der Noten wird mittels Boxplots dargestellt mit Median (horizontale Linie), Mittelwert (Raute) und Interquartilenbereichen (Länge der Boxen). Die vertikalen Linien (Antennen oder Whisker) zeigen Minimal- und Maximalwerte außer den Ausreißern (Punkte) an.

Alpha, basierend auf den Noten an jeder Station berechnet.

Wir berechneten Intraclass-Korrelationskoeffizienten (ICC), um den Anteil der Notenvarianz auszudrücken, der auf Szenarien oder Prüfer/innen zurückzuführen war. Idealerweise sollte dieser Anteil nahe Null sein. Dazu berechneten wir für jede Station getrennt lineare Regressionsmodelle und nutzten wegen der Abweichung der Residuen eine Bootstrap-Verfahren zur Varianzabschätzung, weil keine Normalverteilung vorlag. Wir verwendeten zwei Prädiktorensets:

1. ein Set mit dummy-Kodierung für Szenarien und Prüfer/innen (siehe Tabelle 2);
2. ein Set mit dem vorgenannten Modell plus der Durchschnittsnote aller Stationen außer der betrachteten Station (siehe Tabelle 3).

Die Noten wurden verwendet, um den Einfluss der Leistungen der Studierenden an allen übrigen Stationen außer der Bezugsstation zu berücksichtigen. Berechnungen wurden mit dem xtreg Befehl in Stata unter Anwendung eines fixed-effects Schätzers ausgeführt. Es gab keine fehlenden Daten für die untersuchten Variablen.

Die Analysen wurden in Stata 13 ausgeführt (Stata Corp., College Station, TX)

Tabelle 2: Einfluss von Station und Prüferinnen/Prüfern auf die Note ohne Berücksichtigung der studentischen Fähigkeiten

OSCE Station	Szenario	Häufigkeit (%)	Anzahl verschiedener Prüfer	Durchschnittsnote für jedes Szenario	Gesamtnote der Station	mildester Prüfer Durchschnittsnote (n = geprüfte Studierende)	strengster Prüfer Durchschnittsnote (n = geprüfte Studierende)	ICC für Szenario	ICC für Prüfer
Bauchschmerz	*Gastroenteritis*	57 (33.73)	10	2.00	2.16	1.46 (15)	3.07 (20)	5.9%	39.8%
	Appendizitis	56 (33.14)		2.33					
	Gallekolik	56 (33.14)		2.14					
Husten	*Bronchitis*	57 (33.73)	7	2.27	2.20	1.75 (29)	2.63 (5)	6.2%	20.5 %
	Pneumonie	58 (34.32)		2.30					
	Asthma	54 (31.95)		2.01					
Brustschmerz	*akutes Koronarsyndrom*	59 (34.91)	6	2.52	2.28	1.98 (84)	2.85 (30)	7.8 %	14.1 %
	costosternales Syndrome (Tietze)	56 (33.14)		2.19					
	Reflux	54 (31.95)		2.12					
Müdigkeit	*Psychische Ursache*	80 (47.34)	8	2.37	2.24	1.56 (30)	2.88 (5)	5.2 %	25.7 %
	somatisce Ursache	89 (52.66)		2.14					
Rückenschmerz	*nicht spezifischer Rückenschmerz*	89 (52.66)	7	2.02	2.16	1.51 (15)	2.79 (50)	5.8 %	30.2 %
	radikulärer Schmerz	80 (47.34)		2.32					

ICC=Intraclass Korrelations- Koefficient

Tabelle 3: Einfluss von Station und Prüferinnen/Prüfern auf die Note nach Berücksichtigung der studentischen Fähigkeiten

OSCE Station	Szenario	Häufigkeit (%)	Anzahl verschiedener Prüfer	Durchschnittsnote für jedes Szenario	Gesamtnote der Station	mildester Prüfer Durchschnittsnote (n = geprüfte Studierende)	strengster Prüfer Durchschnittsnote (n = geprüfte Studierende)	ICC für Szenario	ICC für Prüfer
Bauchschmerz	*Gastroenteritis*	57 (33.73)	10	2.00	2.16	1.46 (15)	3.07 (20)	7.2 %	50.3 %
	Appendizitis	56 (33.14)		2.33					
	Gallekolik	56 (33.14)		2.14					
Husten	*Bronchitis*	57 (33.73)	7	2.27	2.20	1.75 (29)	2.63 (5)	4.2 %	24.6 %
	Pneumonie	58 (34.32)		2.30					
	Asthma	54 (31.95)		2.01					
Brustschmerz	*akutes Koronarsyndrom*	59 (34.91)	6	2.52	2.28	1.98 (84)	2.85 (30)	9.2 %	20.5 %
	costosternales Syndrome (Tietze)	56 (33.14)		2.19					
	Reflux	54 (31.95)		2.12					
Müdigkeit	*Psychische Ursache*	80 (47.34)	8	2.37	2.24	1.56 (30)	2.88 (5)	4.5 %	29.2 %
	somatisce Ursache	89 (52.66)		2.14					
Rückenschmerz	*nicht spezifischer Rückenschmerz*	89 (52.66)	7	2.02	2.16	1.51 (15)	2.79 (50)	5.6 %	28.1 %
	radikulärer Schmerz	80 (47.34)		2.32					

ICC=intraclass correlation coefficient

Ergebnisse

Stationen und Prüfer/innen waren ebenso wie die Szenarien statistisch voneinander unabhängig (siehe Anhang 1 und Anhang 2) Die Interne Konsistenz der OSCE-Prüfung über die 5 Notenstufen für die Stationen war gemäß Cronbach's alpha 0,65 ($CI_{90\ one\ sided}$ 0,59).

Vergleich der Szenarien für jede Station

Abbildung 1 zeigt die Verteilung der Noten getrennt für jedes Szenario innerhalb jeder der fünf Stationen sowie die Verteilung der resultierenden Gesamtnoten pro Station.

Die Gesamtnoten der Stationen lagen durchschnittlich zwischen 2,16 und 2,28. Die Differenz der Durchschnittsnoten der Szenarien einer Station betrug zwischen 0,03 bis 0,40 (siehe Tabelle 2 und Tabelle 3). Den größten Unterschied zwischen den Durchschnittsnoten der Szenarien einer Station beobachteten wir an der Station mit dem Beratungsanlass „Brustschmerz". Hier hatte das lebensbedrohliche Szenario akutes Koronarsyndrom (ACS) eine um 0,4 schlechtere Durchschnittsnote gegenüber dem Szenario gastrointestinaler Reflux. An der Station mit dem Beratungsanlass „Bauchschmerz" wurde eine um 0,3 schlechtere Durchschnittsnote bei Szenario Appendizitis gegenüber dem Szenario Gastroenteritis beobachtet. Die Gesamtnoten für die Beratungsanlässe (Stationen) lagen im Bereich von 1 bis 5.

Einfluss von Szenarien und Prüfern auf die Noten an jeder Station

Die Einflüsse von Szenarien und Prüferinen/Prüfern auf die Noten an jeder Station werden als ICCs berechnet und sind in Tabelle 2 und Tabelle 3 dargestellt. In Tabelle 2 zeigen wir die ICCs ohne Berücksichtigung (Adjustierung) der studentischen Fähigkeiten an den übrigen OSCE-Stationen und in Tabelle 3 die ICCs bei Berücksichtigung (Adjustierung) der Fähigkeit der Studierenden an den übrigen Stationen. Der Einfluss der Szenarien auf die Gesamtnoten der entsprechenden Stationen betrug 5,2% bis 7,8%, wenn keine Berücksichtigung der Durchschnittsnoten der Studierenden an den übrigen Stationen erfolgte. Bei Berücksichtigung der Fertigkeiten der Studierenden an den übrigen Stationen betrug der Einfluss der Szenarien 4,2% bis 9,2%. Bei Betrachtung der größten Differenz zwischen den Durchschnittsnoten der Szenarien einer Station konnte der größte Einfluss eines Szenarios auf die Gesamtnote an der Station mit dem Beratungsanlass „Brustschmerz" festgestellt werden.

Die Anzahl der Prüfer/innen an jeder Station bewegte sich zwischen 6 und 10 über die drei Prüfungstage. Der nicht adjustierte Einfluss der Prüfer/innen auf die Gesamtnoten an einer Station schwankte zwischen 14,1% und 39,8% ohne Berücksichtigung der Durchschnittsnoten der Studierenden an den übrigen Stationen. Er betrug zwischen 20,5% und 50,3%, wenn die studentischen Fähigkeiten an den übrigen Stationen berücksichtigt wurden (Adjustierung). Der größte Prüfereffekt wurde an der Station mit dem Beratungsanlass „Bauchschmerz" gesehen.

Diskussion

Zusammenfassung der Hauptergebnisse

Insgesamt nahmen 169 Studierende im 2. klinischen Jahr und 23 Prüferinnen und Prüfer an der MS-OSCE teil. Die Differenz der Durchschnittsnoten der Szenarien eines Beratungsanlasses (Station) betrug 0,03 bis 0.4 auf einer Notenskala von 1-5. Der Einfluss der Szenarien einer Station auf deren Gesamtbenotung erklärte 4% bis 9% der Notenschwankungen. Bei Berücksichtigung der studentischen Fähigkeiten an den übrigen Stationen war der Einfluss der Prüferinnen und Prüfer an einer Station für 20% bis 50% der Gesamtbenotungsschwankungen der Station verantwortlich.

Bedeutung der Ergebnisse

Wir beobachteten Notenunterschiede von 0,03 bis 0,4 zwischen den Szenarien der gleichen Station auf einer Notenscala von 1-5 (siehe Abbildung 1). Auch wenn die Checklist-Items für jedes Szenario eines Beratungsanlasses die gleiche Vorgehensroutine abdeckten, sollten die Bewertungen eigentlich nicht von der Gefährlichkeit der zugrundeliegenden Diagnose beeinflusst worden sein, da wir verlangten, dass die Studierenden alle Möglichkeiten in Erwägung zogen. Wurde die Diagnose eines potentiell lebensgefährlichen Szenarios verpasst oder dessen Management misslang, scheint das zu einer schlechteren Benotung geführt zu haben. Das war der Fall, wenn ein akutes Koronarsyndrom, eine Appendizitis oder eine Pneumonie vorlag, während ähnliche Fehler bei einem eher gutartigen Szenario wie costosternalem Syndrom, Gastroenteritis oder Bronchitis nicht mit einer schlechteren Benotung einherging.

Es besteht kein Konsens darüber, ab wann eine Notendifferenz als bedeutsam für die Objektivität einer Prüfung gewertet werden soll. Wir schätzen die Bedeutung der beobachteten Notendifferenzen als eher gering bis moderat ein. Verglichen mit dem Ausmaß des Einflusses verschiedener Prüfer/innen auf die Noten einer Station erscheint der Effekt unterschiedlicher Szenarien klein aber noch relevant zu sein. Der Einfluss der Prüfer/innen auf die Noten der Studierenden war unabhängig von den Szenarien und von den Fähigkeiten der Studierenden. Aber der Unterschied in der Durchschnittsnote zwischen dem nachsichtigstem und dem strengsten Prüfer lag bei über 1 Notenstufe auf der 5-teiligen Notenskala, was eine schlechte Inter-Rater-Reliabilität (Zuverlässigkeit der Einschätzung durch verschiedene Prüfer/innen an der gleichen Station) wahrscheinlich macht. Daher scheint eine Prüfer/innenschulung (Vermittlung einheitlicher Bewertungskriterien für ein beobachtetes Verhalten) weitaus bedeutender zu sein als eine Anpassung der unterschiedlichen Schwierigkeiten der Szenarien. Wilkinson et al. [23] zeigten, „dass Prüferfaktoren substantiell stärker zur Objektivität einer OSCE-Prüfung beitragen, als es Bewertungsbögen oder Checklisten tun". Über Inter-Rater-Reliabilität bei OSCE-Prüfungen wurde nur wenig veröffentlicht und sie variiert je nach OSCE-Aufbau, eingesetzten Prüfungsinstrumenten (global rating/ checklist rating) und den Prüfungsbedingungen (direkte Beobachtung der Prüfungssituation/ Beobachtung eines Prüfungsvideos) [20], [24], [25]. Hatala et al. [26] pilotierten eine OSCE-Prüfung, bei der 2 Stationen in 3 aufeinanderfolgende Prüfungssequenzen zu 10 Minuten unterteilt wurden, wobei verschiedene Aspekte eines Problems in der Inneren Medizin abgedeckt wurden. Sie beobachteten eine Inter-Rater-Reliabilität zwischen 0,63 bis 0,91 bei 2 Prüfern für jedes Szenario. Brennan et al [16] berichten, dass, obwohl sich die Spanne der Notengebung änderte, wenn die Prüfer/innen an den OSCE-Stationen wechselten, die Reliabilität der Prüfung und die Ergebnisse der Prüflinge nicht beeinträchtigt wurden (die Gesamtzahl der Prüfer/innen an einer Station wird allerdings nicht mitgeteilt).

Wegen beschränkter finanzieller Ressourcen konnten wir - wie auch viele andere medizinische Fakultäten – es uns nicht leisten, jede OSCE mit zwei Prüfern gleichzeitig zu besetzen.

Intensiveres Training von Prüferinnen/Prüfern und SPs [4] sowie eine noch sorgfältigere Entwicklung von Checklisten sind mögliche Maßnahmen, um den Effekt von Prüferinnen/Prüfern auf die Notengebung zu reduzieren und so eine bessere Inter-Rater-Reliabilität zu erreichen. Die Annahme, dass ein intensiveres Prüfer/innen -Training die Inter-Rater-Reliabilität erhöht, kann nicht generalisiert werden [27], [28]. In welchem Maße Unfairness und Fehlen von Reliabilität akzeptabel sind und wie sehr sich der Einfluss der Prüfer/innen reduzieren lässt, wird noch diskutiert [29].

Wir nehmen nicht an, dass das MS-OSCE Prüfungsformat den Informationsaustausch unter Studierenden reduziert hat, aber wir nehmen an, dass der Wechsel zu MS-OSCE dazu geführt hat, dass Anamnese und klinische Untersuchung an den Stationen während der 3 Prüfungstage gründlicher und weniger hastig ausgeführt werden. Objektive Daten, um diese Annahme zu stützen, haben wir allerdings nicht.

Stärken und Schwächen

Dies ist nach unserer Kenntnis der erste Bericht über eine Multiple Scenario-OSCE-Prüfung. Wir berechneten adjustiert für die Fähigkeiten der Studierenden den Einfluss

von multiplen Szenarien und von Prüferinnen und Prüfern auf die Prüfungsnoten bei einer MS-OSCE-Prüfung. Aufgrund limitierter Ressourcen konnten wir keinen Inter-Rater-Korrelationen für die Checklisten ermitteln und nur ein minimales Prüfer/innentraining realisieren. Diese Situation dürfte bei den meisten Medizinischen Fakultäten, die studentische Fähigkeiten mit einer OSCE -prüfen, ähnlich sein. Zwischen der Bewertung mittels Checkliste und der Globalbewertung zeigte sich ein Korrelation von 0,6 bis 0,8, was für eine kongruente Bewertung von kommunikativen und klinischen Fertigkeiten spricht (Ergebnisse nicht dargestellt). Wir können nicht ausschließen, dass eine unterschiedliche Genauigkeit der Darstellung des gleichen Szenarios durch verschiedene SPs während der 3-tägigen Prüfung einen Einfluss auf die Bewertung hatte. Eine Adjustierung der SPs haben wir nicht durchgeführt. Auch Gender-Effekte, die einen Einfluss auf die Notengebung haben können, wurden von uns nicht berücksichtigt [29], [30], [31]. Weibliche und männliche SPs wechselten an einigen Stationen, was die Performance der Studierenden an der „Brustschmerz"-Station und der Station mit dem Beratungsanlass „akuter Husten", an denen die Auskultation des Thorax als mögliche klinische Untersuchung in Frage kam, beeinflusst haben kann. Unsere MS-OSCE-Prüfung mit nur fünf Stationen für die Beurteilung einer/eines Studierenden ist relativ kurz da für eine reliable Prüfung mindestens 10 Stationen eingerichtet werden sollten [32], [33]. Zehn Minuten pro Station sind eine akzeptable Zeitspanne [34], [35] und sogar für Abschluss/Aufnahmeprüfungen mit höchsten Anforderungen werden nur 15 Minuten für eine interaktive Aufgabenstellung verlangt [36]. Wir haben eine gute interne Konsistenz (Cronbach's alpha: 0,65) über alle Stationen, verglichen mit anderen Veröffentlichungen [32].

Auch wenn es möglich ist, nach der Prüfung die individuelle Note einer/eines Studierenden für Unterschiede innerhalb der Szenarien und Unterschiede zwischen den Prüfern mit einem Korrekturfaktor zu adjustieren, haben wir das nicht getan. Die Berechnung von Korrekturfaktoren nach jeder Prüfung würde Ressourcen erfordern, die uns derzeit nicht zur Verfügung stehen.

Die Untersuchung der Validität einer MS-OSCE-Prüfung ist nicht Gegenstand unserer Veröffentlichung. Van der Vleuten and Schuwirth [7] stellen fest, dass Schlüsselparameter für die Validität von Kompetenzeinschätzungen die Authentizität der gezeigten Leistung (Performance) und die Einbeziehung von professionellen Kompetenzen sind. Die MS-OSCE-Prüfung zielt auf die Authentizität der Performance der Studierenden ab, indem sie alternierend mehrere Szenarien für den Beratungsanlass einer Station anbietet, um so den Einfluss von weitergegebenen Informationen (Weitersagen) auf den Umgang mit der Aufgabenstellung der Studierenden zu reduzieren. Die Inhaltsvalidität der Prüfung wurde durch Reviewing aller MS-OSCE-Stationen durch ein Team erfahrener Lehrärztinnen und Lehrärzte angestrebt. Durch die Darstellung klinischer Szenarien mit SPs an jeder der fünf MS-OSCE-Stationen und durch den Einsatz standardisierter Checklisten und einem validierten Globalbewertungsinstrument sollte die Augenscheinvalidität der MS-OSCE der einer traditionellen OSCE mit 5 Stationen gleichen.

Schlussfolgerung

Der Einfluss verschiedener Szenarien auf die Examensnote für das Management eines Beratungsanlasses in der Allgemeinmedizin war gering im Vergleich zum Einfluss der Prüferinnen und Prüfer. Um Objektivität und Fairness einer OSCE-Prüfung zu gewährleisten ist es bedeutsamer, die Inter-Rater-Reliabilität zu verbessern, als alle Studierenden mit dem gleichen Szenario zu prüfen.

Liste der Abkürzungen

ACS: Akutes Koronarsyndrom (acute coronary syndrome)
BGR: Berliner Global Rating Instrument
CI: Konfidenzintervall (confidence interval)
CR: Checklistenprüfung (checklist rating)
OGR: globale Gesamtbeurteilung (overall global rating)
ICC: Intraclass-Korrelationskoeffizienten (intraclass correlation coefficient)
MS-OSCE: Multiple Scenario Objective Structured Clinical Examination
OSCE: Objective Structured Clinical Examination
SP: Simulationspatientin/Simulationspatient

Danksagungen

Wir danken den Lehrärzten Francis Baudet, Gisela Greschniok, Heinz Hammermayer, Thomas Hannemann, Mathias Herberg, Gero Kärst, Andreas Krüger, Barbara Krüger, Annika Matz, Hans-Diether Seiboth, Thomas Richter, Claudia Runge, Carmina Spreemann, Antje Theurer, Renate Tilchner, Rüdiger Titze, Arne Wasmuth, Christine Wendt, Arno Wilfert.

Erhältlichkeit weiterer Daten

Weitere Daten können in begründeten Fällen vom Korrespondenzautor zur Verfügung gestellt werden.

Anteilige Mitwirkung der Autoren

JS und JFC hatten die Idee zur MS-OSCE-Prüfung. Szenarien und Prüfungsmaterialien wurden entwickelt und pilotiert von JS, CR, GW, AA, FL, AH, JFC. CR, JS, FL und GW trainierten die Simulationspatientinnen und Simulationspatienten, AH war für die Datenverwaltung verantwortlich, COS führte die statistische Auswertung durch. JS und JFC schrieben den Textentwurf, der von allen Autoren geprüft und freigegeben wurde.

Interessenkonflikt

Die Autoren erklären, dass sie keine Interessenkonflikte im Zusammenhang mit diesem Artikel haben.

Anhänge

Verfügbar unter
http://www.egms.de/en/journals/zma/2019-36/zma001234.shtml

1. Anhang_1.pdf (238 KB)
   Unabhängigkeit von Stationen und Prüferinnen/Prüfern. Information zu Anhang 1: Zur Bewertung der Unabhängigkeit wurde Chi² -Test oder Fisher's Exact Test berechnet.
2. Anhang_2.pdf (251 KB)
   Unabhängigkeit der Szenarien. Information zu Anhang 2: Zur Bewertung der Unabhängigkeit wurde Chi² -Test oder Fisher's Exact Test berechnet.

Literatur

1. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J. 1975;1(5955):447-451. DOI: 10.1136/bmj.1.5955.447
2. Vu NV, Barrows HS. Use of Standardized Patients in Clinical Assessments: Recent Developments and Measurement Findings. Educ Res. 1994;23:23-30. DOI: 10.3102/0013189X023003023
3. Patrício MF, Julião M, Fareleira F, Carneiro AV. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? Med Teach. 2013;35(6):503-514. DOI: 10.3109/0142159X.2013.774330
4. Baig LA, Beran TN, Vallevand A, Baig ZA, Monroy-Cuadros M. Accuracy of portrayal by standardized patients: results from four OSCE stations conducted for high stakes examinations. BMC Med Educ. 2014;14:97. DOI: 10.1186/1472-6920-14-97
5. Van der Vleuten CP, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. Med Educ. 1991;25(2):110-118. DOI: 10.1111/j.1365-2923.1991.tb00036.x
6. Furman GE, Smee S, Wilson C. Quality assurance best practices for simulation-based examinations. Simul Healthc. 2010;5(4):226-231. DOI: 10.1097/SIH.0b013e3181da5c93
7. van der Vleuten CP, Schuwirth LW. Assessing professional competence. From methods to programmes. Med Educ. 2005;39(3):309-317. DOI: 10.1111/j.1365-2929.2005.02094.x
8. Parks R, Warren PM, Boyd KM, Cameron H, Cumming A, Lloyd-Jones G. The Objective Structured Clinical Examination and student collusion: marks do not tell the whole truth. J Med Ethics. 2006;32(12):734-738. DOI: 10.1136/jme.2005.015446
9. Colliver JA, Barrows HS, Vu NV, Verhulst SJ, Mast TA, Travis TA. Test security in examinations that use standardized-patient cases at one medical school. Acad Med. 1991;66(5):279-282. DOI: 10.1097/00001888-199105000-00011
10. Colliver JA, Travis TA, Robbs RS, Barnhart AJ, Shirar LE, Vu NV. Test security in standardized-patient examinations: analysis with scores on working diagnosis and final diagnosis. Acad Med. 1992;67(10):S7-S9. DOI: 10.1097/00001888-199210000-00022
11. Harden RM, Lilley P, Patricio M. The definitive guide to the OSCE: The Objective Structured Clinical Examination as a performance assessment. Edinburgh, New York: Elsevier; 2016.
12. Kennedy G, Gray K, Tse J. 'Net Generation' medical students: technological experiences of pre-clinical and clinical students. Med Teach. 2008;30(1):10-16. DOI: 10.1080/01421590701798737
13. Pander T, Pinilla S, Dimitriadis K, Fischer MR. The use of Facebook in medical education - a literature review. GMS Z Med Ausbild. 2014;31(3):Doc33. DOI: 10.3205/zma000925
14. Rutala PJ. Sharing of Information by Students in an Objective Structured Clinical Examination. Arch Intern Med. 1991;151(3):541. DOI: 10.1001/archinte.1991.00400030089016
15. Wilkinson TJ, Fontaine S, Egan T. Was a breach of examination security unfair in an objective structured clinical examination? A critical incident. Med Teach. 2003;25(1):42-46. DOI: 10.1080/0142159021000061413
16. Brennan PA, Croke DT, Reed M, Smith L, Munro E, Foulkes J, Arnett R. Does Changing Examiner Stations During UK Postgraduate Surgery Objective Structured Clinical Examinations Influence Examination Reliability and Candidates' Scores? J Surg Educ. 2016;73(4):616-623. DOI: 10.1016/j.jsurg.2016.01.010
17. Chenot JF. Undergraduate medical education in Germany. GMS Ger Med Sic. 2009;7:Doc02. DOI: 10.3205/000061
18. Scheffer S. Validierung des "Berliner Global Rating" (BGR). Ein Instrument zur Prüfung kommunikativer Kompetenzen Medizinstudierender im Rahmen klinisch-praktischer Prüfungen (OSCE) [An instrument for assessing communicative competencies of medical students within the frame of testing clinical skills]. Berlin: Charité - Universitätsmedizin Berlin, Medizinische Fakultät; 2009. Zugänglich unter/available from: http://nbn-resolving.de/urn:nbn:de:kobv:188-fudissthesis000000010951-7
19. Regehr G, Freeman R, Robb A, Missiha N, Heisey R. OSCE performance evaluations made by standardized patients: comparing checklist and global rating scores. Acad Med. 1999;74(10 Suppl):S135-S137. DOI: 10.1097/00001888-199910000-00064
20. Ilgen JS, Ma IWY, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. Med Educ. 2015;49(2):161-173. DOI: 10.1111/medu.12621
21. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Med Educ. 2003;37(11):1012-1016. DOI: 10.1046/j.1365-2923.2003.01674.x
22. Hunter DM, Jones RM, Randhawa BS. The use of holistic versus analytic scoring for large-scale assessment of writing. Can J Prog Eval. 1996;11:61-85.
23. Wilkinson TJ, Frampton CM, Thompson-Fawcett M, Egan T. Objectivity in Objective Structured Clinical Examinations: Checklists Are No Substitute for Examiner Commitment. Acad Med. 2003;78(2):219-223. DOI: 10.1097/00001888-200302000-00021
24. Kim J, Neilipovitz D, Cardinal P, Chiu M. A comparison of global rating scale and checklist scores in the validation of an evaluation tool to assess performance in the resuscitation of critically ill patients during simulated emergencies (abbreviated as "CRM simulator study IB"). Simul Healthc. 2009;4(1):6-16. DOI: 10.1097/SIH.0b013e3181880472
25. Malau-Aduli BS, Mulcahy S, Warnecke E, Otahal P, Teague PA, Turner R, van der Vleuten C. Inter-Rater Reliability: Comparison of Checklist and Global Scoring for OSCEs. Creat Educ. 2012; 03:937-942. DOI: 10.4236/ce.2012.326142

26. Hatala R, Marr S, Cuncic C, Bacchus CM. Modification of an OSCE format to enhance patient continuity in a high-stakes assessment of clinical performance. BMC Med Educ. 2011;11:23. DOI: 10.1186/1472-6920-11-23

27. Weitz G, Vinzentius C, Twesten C, Lehnert H, Bonnemeier H, König IR. Effects of a rater training on rating accuracy in a physical examination skills assessment. GMS Z Med Ausbild. 2014;31(4):Doc41. DOI: 10.3205/zma000933

28. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores. A randomized, controlled trial. J Gen Int Med. 2009;24(1):74-79. DOI: 10.1007/s11606-008-0842-3

29. Schleicher I, Leitner K, Juenger J, Moeltner A, Ruesseler M, Bender B, Sterz J, Schuettler KF, Koenig S, Kreuder JG. Examiner effect on the objective structured clinical exam - a study at five medical schools. BMC Med Educ. 2017;17(1):71. DOI: 10.1186/s12909-017-0908-1

30. Mortsiefer A, Karger A, Rotthoff T, Raski B, Pentzek M. Examiner characteristics and interrater reliability in a communication OSCE. Pat Educ Couns. 2017;100(6):1230-1234. DOI: 10.1016/j.pec.2017.01.013

31. Carson JA, Peets A, Grant V, McLaughlin K. The effect of gender interactions on students' physical examination ratings in objective structured clinical examination stations. Acad Med. 2010;85(11):1772-1776. DOI: 10.1097/ACM.0b013e3181f52ef8

32. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Med Educ. 2011;45(12):1181-1189. DOI: 10.1111/j.1365-2923.2011.04075.x

33. Nikendei C, Jünger J. OSCE - hands on instructions for the implementation of an objective structured clinical examination. GMS Z Med Ausbild. 2006;23(3):Doc47. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2006-23/zma000266.shtml

34. Reznick RK, Blackmore D, Dauphinee WD, Rothman AI, Smee S. Large-scale High-stakes Testing with an OSCE: Report from the Medical Council of Canada. Acad Med. 1996;71(1 Suppl):S19-S21. DOI: 10.1097/00001888-199601000-00031

35. Hamann C, Volkan K, Fishman MB, Silvestri RC, Simon SR, Fletcher SW. How well do second-year students learn physical diagnosis? Observational study of an objective structured clinical examination (OSCE). BMC Med Educ. 2002;2:1-11. DOI: 10.1186/1472-6920-2-1

36. Chambers KA, Boulet JR, Gary NE. The management of patient encounter time in a high-stakes assessment using standardized patients. Med Educ. 2000;34(10):813-817. DOI: 10.1046/j.1365-2923.2000.00752.x

Korrespondenzadressen:

Johannes Spanke
University Medicine Greifswald, Institute for Community Medicine, Department of General Practice and Family Medicine, Fleischmannstr. 6, 17475 Greifswald, Deutschland
johannes.spanke@uni-greifswald.de
Christina Raus
University Medicine Greifswald, Institute for Community Medicine, Department of General Practice and Family Medicine, Fleischmannstr. 6, 17475 Greifswald, Deutschland

Bitte zitieren als
Spanke J, Raus C, Haase A, Angelow A, Ludwig F, Weckmann G, Schmidt CO, Chenot JF. Fairness and objectivity of a multiple scenario objective structured clinical examination. GMS J Med Educ. 2019;36(3):Doc26.
DOI: 10.3205/zma001234, URN: urn:nbn:de:0183-zma0012343

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/zma/2019-36/zma001234.shtml

Eingereicht: *23.05.2018*
Überarbeitet: *11.11.2018*
Angenommen: *13.02.2019*
Veröffentlicht: *16.05.2019*