What works best in a general practice specific OSCE for medical students: mini-CEX or content-related checklists?

Patrick Giemsa, Clara Wübbolding, Martin R. Fischer, Tanja Graupe, Anja Härtl, Christine Lenz, Linda Sanftenberg, Jörg Schelling, Katrin Schüttpelz-Brauns, Claudia Kiessling

Angaben zur Veröffentlichung / Publication details:

Giemsa, Patrick, Clara Wübbolding, Martin R. Fischer, Tanja Graupe, Anja Härtl, Christine Lenz, Linda Sanftenberg, Jörg Schelling, Katrin Schüttpelz-Brauns, and Claudia Kiessling. 2020. "What works best in a general practice specific OSCE for medical students: mini-CEX or content-related checklists?" *Medical Teacher* 42 (5): 578–84. https://doi.org/10.1080/0142159x.2020.1721449.

What works best in a general practice specific OSCE for medical students: Mini-CEX or content-related checklists?

Patrick Giemsa[a], Clara Wübbolding[b], Martin R. Fischer[b], Tanja Graupe[c,b], Anja Härtl[d,b], Christine Lenz[e], Linda Sanftenberg[e], Jörg Schelling[e], Katrin Schüttpelz-Brauns[f] and Claudia Kiessling[a,b]

[a]Lehrstuhl für die Ausbildung Personaler und Interpersonaler Kompetenzen im Gesundheitswesen, Fakultät für Gesundheit, Universität Witten/Herdecke, Witten, Germany; [b]Institut für Didaktik und Ausbildungsforschung in der Medizin, Klinikum der Universität München, Munich, Germany; [c]MeCuM-Mentor, Klinikum der Universität München, Munich, Germany; [d]Department for Medical Education and Educational Research, Faculty of Medicine, University of Augsburg, Augsburg, Germany; [e]Institut für Allgemeinmedizin, Klinikum der Universität München, Munich, Germany; [f]Department of Undergraduate Education and Educational Development, Medical Faculty, Mannheim at Heidelberg University, Mannheim, Germany

ABSTRACT

Aim: To develop and pilot a General Practice (GPr) OSCE assessing medical students dealing with patient encounters, which are typical for GPr and to compare different measurement instruments (global ratings, content-specific checklists).

Methods: A blueprint based on Entrusted Professional Activities was used to develop prototypical OSCE stations. Four stations were tested with voluntary medical students. Students were videotaped and assessed with self-developed content-specific checklists, a global rating for communication skills, and mini-CEX. Results were compared according to students' phases of studies.

Results: All three measurements were able to discriminate between clinical and pre-clinical students. Clearest results were achieved by using mini-CEX. Content-specific checklists were not able to differentiate between those groups for the more difficult stations. Inter-station reliability for the global ratings was sufficient for high-stakes exams. Students enjoyed the OSCE-setting simulating GPr consultation hours. They would prefer feedback from GPs after the OSCE and from simulated patients after each encounter.

Discussion and conclusion: Although the OSCE was short, results indicate advantages for using a global rating instead of checklists. Further research should include validating these results with a larger group of students and to find the threshold during the phases of education for switching from checklists to global ratings.

KEYWORDS

Ambulatory medicine; OSCE; sychometrics; undergraduate; medicine

Introduction

Many activities have been implemented in recent years to enhance the visibility and attractiveness of General Practice (GPr) in Germany. Background for these activities is a major concern that primary care services especially outside large cities might not be sufficiently available. Demographic development with an increase of elderly patients and chronic patients in combination with a decrease of practicing general practitioners (GPs) in rural areas has led to this concern (Kopetsch 2010). Additionally, a marked shift away from GPr to other medical specialties has taken place in the health care systems of most Western countries, especially in those with competition-based health care systems (Hasler et al. 2008). Working conditions of GPs does not seem to be very attractive to young doctors. Different studies mentioned the high workload, irregular working hours, as well as the low income and occupational prestige in comparison to other medical subjects as the main reasons for this development (Whitcomb and Cohen 2004; Walker 2006; Buddeberg-Fischer et al. 2008).

Practice points

- As for all OSCEs, special care should be taken in the selection of representative stations with various difficulties. Self-developed EPAs used as a blueprint can help with the selection.
- Good inter-station reliability can be achieved with as few as 4 stations using one rater and video analysis.
- When choosing between global ratings and behavioral checklists, former are more sensitive in discriminating between clinical and pre-clinical students. This is especially true when focusing on clinical competence and the ability to adapt to changing circumstances that require communication skills and empathy.
- A naturalistic OSCE setting, where SP rather than students rotate from station to station is perceived to adequately simulate a GPr consultation and can be used in a formative and summative assessment.

CONTACT Claudia Kiessling claudia.kiessling@uni-wh.de Lehrstuhl fuer die Ausbildung Personaler und Interpersonaler, Kompetenzen im Gesundheitswesen, Universität Witten/Herdecke, Fakultät für Gesundheit, Pferdebachstr. 11, 58455 Witten, Germany.

The study was conducted at the University Hospital of LMU Munich, Germany

For instance, of all practicing physicians in Switzerland in 2006, only 13.6% were GPs, and of all newly qualified physicians in 2006, only 13.4% specialized in GPr. In Germany, the percentage of GPs has also declined, and only half of the physicians who qualified as GPs in the last years opened private practices. Overall, the shortage of physicians providing primary health care has not only being noticed in rural regions but is also starting to be a problem in metropolitan areas (Buddeberg-Fischer et al. 2007).

In Germany, GPr has been strengthened in undergraduate medical education by including lectures, compulsory courses and attachments during the clinical years of medical studies (Gensichen and Gerlach 2003). Especially the optional training in GPr within the last year of studies was an important implementation in the undergraduate medical education in year 2003. It is under discussion to transform this voluntary rotation into a compulsory one. Numerous organisational and administrative challenges have to be considered in this setting and recruited practices as well as university departments of GPr should expect an additional workload. However, students might profit immensely from interactive learning experiences, which go far beyond usual rotations in hospital-based clinical departments (Gulich 2005). These activities allow students to gain an insight into how GPs work and with what kind of problems they usually have to deal with. German associations for GPr were able to establish quality standards for teaching and training in GPr outside university hospitals, e.g. by defining criteria for GPs involved in teaching and training, establishing learning outcomes for courses and electives, developing a logbook for students in their final year working in GPr (DEGAM e.V. 2013).

However, a field of improvement is formative and summative assessment of students dealing with the specific problems, which patients present when they come to a GP. According to the principles of constructive alignment (Biggs 2003), assessment is an integral part of a well-designed learning and teaching activity. It allows to clearly address intended outcomes to learners from clearly specified objectives to well informed feedback on learners' performance and knowledge. A well-designed assessment format can therefore highlight the importance of the intended outcome, in this case the role of GPr in medical education.

The integration of Entrusted Professional Activities (EPAs) in training and formative assessment may close the gap between the theory of competency-based training and patient-centred practice (Peters et al. 2017; Ten Cate 2017; Ten Cate et al. 2018). EPAs may also function as a blueprint for assessment programs. Different assessment formats have been described to assess medical students including oral and written examinations, Objective Structured Clinical Examination (OSCE) logbooks, and workplace-based assessment (Norcini and Burch 2007; Swayamprakasam et al. 2014). By using one of these formats in a valid and reliable way, attention must be paid to test content, test design, and implementation factors, especially when the results will be used for high-stakes decision making. In recent years, OSCEs have become one of the most popular assessment formats to test clinical competence in undergraduate medical education. While questions remain around the application of OSCE testing, there are also both known and hidden benefits to students, faculty, and organizations that use OSCEs (Townsend et al. 2001; Wilkinson et al. 2003; Turner and Dankoski 2008). Educators subjectively believe in high-fidelity assessment, and students and educators generally feel positive about this type of performance testing. In a publication, reporting student, curricular, and faculty development outcomes after nine years of OSCE testing (Duerson et al. 2000), student performance improved, small-group teaching sessions were standardized, and faculty received feedback that improved instruction and enthusiasm for teaching physical exam skills. Also, the use of OSCEs for evaluation reinforces the patient-centered nature of medical practice, often provides timely and specific feedback on clinical performance (Barrows 1993). While most OSCEs use behavioural checklists when assessing student's abilities in demonstrating single procedures (e.g. vein puncture, history taking), there is still a debate about which instruments should be used to evaluate competencies in more complex situations. Many studies have compared content-specific checklists to broader global ratings of clinical competence, but most only focus on a graduate population or comparing graduate and undergraduate students (Cunnington et al. 1996; Ilgen et al. 2015). The aim of this study was to develop and pilot a GPr-specific OSCE and to explore which kind of instruments to assess students' performance presented in the OSCE would discriminate best between preclinical students and clinical students, thus comparing the two main phases of undergraduate studies in Germany.

Methods

Setting of the study

Over 900 medical students start their medical course of undergraduate training at the Ludwig-Maximilians-University (LMU) in Munich, Germany. After two years when the students have passed the preclinical part of the national licensing examination, 60% of the cohort continues their clinical studies at LMU whereas the remaining 40% of the cohort continue their training at the Technical University. During the clinical phase of studies, students attend lectures and seminars dealing with GPr specific health problems and a two-week attachment at a GP (during year four or five). In their final year of studies, only a small group of students (around 5%) decide to spend a four-month-rotation in GPr. Assessment during the clinical phase of studies includes a written case presentation (during the 2-week attachment) and a logbook during the final year. There is no formative or summative assessment to evaluate students' performance based on observation in direct contact with simulated patients (SP) or real patients. This course of studies is quite typical for medical schools in German-speaking countries.

Study design

We piloted a GPr-specific OSCE (GPr-OSCE) to test the setting and different measurement instruments. We were interested in instruments, which can be used in workplace-based assessments as well as in OSCEs. The GPr-OSCE included patient problems, which were typical for primary

care on the one hand and a simulation of a GP consultation setting on the other hand. To define different patient problems and scenarios, we used self-developed EPAs as a blueprint with the following categories: (1) dealing with patients with an acute health problem, (2) dealing with patients with chronic health problems, (3) emergency situations in primary care, (4) home visits, (5) prevention, (6) intersection in health care, and (7) ambulant palliative care. In this first pilot, we focussed on four patient problems, which were perceived as most common and relevant for students in a GPr setting.

Usually, in an OSCE, students rotate through a circuit of different stations by walking from room to room. Assessors and simulated patients stay in one room and one student after another enter the room to fulfil the requirements. In our GPr-OSCE, students stay in one room (like in a consulting room) and one SP after another enter the room and present their problem. Therefore, SP walk from room to room and rotate through the circuit. We used four stations with the following patient problems based on real patient histories:

- 53-year-old female patient with subacute headache. Background is a psychosocial problem because her partner suffered a stroke a few weeks ago and the housing situation is now unclear ('headache-station').
- 35-year-old female patient with acute but harmless extrasystolic heart beats. She also wants to inform herself about a regular health check-up for 35-year-old women ('health-check-up-station').
- 57-year-old male patient after hospitalisation due to acute chest pain. He received five new drugs in the hospital to reduce his risk factors, which he was unaware of before the hospital stay. He is irritated and does not feel well informed ('Intersection-station').
- 75-year-old male patient with diabetes mellitus II. His blood sugar values are insufficient. He wants another type of insulin form another company, an insulin his neighbour uses ('Diabetes-station').

Due to the setting of the OSCE with video-analysis and a restricted time-limit of 10 minutes per encounter, we excluded physical examination as part of the requirements. For each station, we used the following measurement instruments to assess students' abilities in gathering information, sharing information, clinical reasoning, setting up an action plan, and communication skills:

- Mini-CEX (6 items) 9-point scale (1 = not sufficient, 9 = outstanding). The mini-CEX was designed to test different aspects of clinical performance like history taking, professionalism, clinical judgment, efficiency, and overall clinical care (Norcini et al. 2003).
- Berlin Global Rating (BGR, Scheffer et al. (2008)) for assessing communication skills i.e. empathy, structure, verbal and nonverbal communication (4 items, 5-point-scale with verbal anchors for each item, 1 = incompetent; 5 = competent), which is based on the analytical global rating published by Hodges et al. (2002).
- A content-specific checklist for each station (3-point scale: 2 = fulfilled, 1 = partly fulfilled, 0 = not fulfilled). Checklists for all stations were developed by a medical doctor according to current guidelines. All checklists were reviewed by at least two GPs.

To test the GPr-OSCE, we used an experimental design with the phase of study as our independent variable and test scores as our dependent variable. We asked medical students from the first to the last years of studies to participate in our study. Participation was voluntary and reimbursed by 20 Euros. We invited students from all years of studies to achieve a broad variance in our results. Students had 20 (clinical students) to 30 minutes (preclinical students) time to prepare the stations. Students received simulated patient records and medical background information to prevent excessive demands for especially the preclinical students (e.g. information about drugs, information about diabetes mellitus treatment). Due to this being a pilot study and organisational reasons (time schedule), total testing time was limited to 40 minutes. During the assessment, SPs handed out additional material if necessary, like a blood sugar diary (Diabetes station) or a discharge letter (Intersection station).

As students were videotaped during each SP consultation, ratings of students' performance were executed by analysing these videos. For each station and measurement instrument, a rater training was conducted with three to five raters depending on the instrument: (1) one or two medical students and doctorate candidates, (2) a medical doctor and expert in assessment, (3) a GP, and (4) an educationalist. Each training lasted about two hours watching, rating, and discussing one or two videos per station. Then, each assessor rated 10 stations in parallel. Interrater reliability between the assessment expert and the other raters within the CEX and BGR was calculated using ICC. For the checklists, we used Cohen's kappa as a measurement for interrater reliability. All results between the expert and our final rater (one of the doctorate candidates) were sufficiently high (on average $r = 0.88$) The final rater then rated all videos from the GPr-OSCE according to the coding protocol.

We used R v.3.5.1 to analyze the data, setting our level of significance to 5%. Inter-station reliability of each scale was calculated using Cronbach's alpha. Differences between preclinical and clinical students were evaluated by a multiple analysis of variance (MANOVA), followed by separate ANOVA for each independent variable hypothesizing that students in the clinical years would score higher than students in the preclinical years.

After completing the OSCE, students were asked to fill out a questionnaire to evaluate the OSCE setting and stations. They were also asked to evaluate different ways of receiving feedback: feedback from SP or GPs, feedback directly after each encounter, after the whole OSCE or based on video analysis. Each feedback method was evaluated separately on a seven-point scale (1 = totally disagree, 5 = totally agree).

Results

Study population

Eighty-eight students participated in our pilot GPr-OSCE. Of these, 33 studied in the preclinical years (year 1–2), 51 studied in the clinical years (year 3–5) and four students

were in the last year of studies (5%). Of the students, 65 were female (74%). Mean age was 23. Most students were German native speakers (81%).

Internal consistency and correlations

For our content-specific checklists, alpha was $\alpha = 0.57$. Mean score was $M = 54\%$ (SD = 9.6), Standard error of measurement was $SEM = 6.3$. Inter-station reliability for the Mini-CEX was $\alpha = .84$ with a mean score of $M = 55\%$ (SD = 14.7), and SEM = 5.8. Cronbach's alpha for the BGR was $\alpha = .88$. Mean score was $M = 62\%$ (SD = 13.6), SEM was SEM = 4.7. Of the four stations, the intersection station was the most difficult one. The easiest station was the headache station. Looking at the checklist items, students had in particular difficulties in the following tasks: stating a tentative diagnosis and exploring a full pain history, exploring the psychosocial history and patient perspective, exploring risk factors, explaining findings and medication, and make an action plan. Looking at the Mini-CEX items, students had most difficulties in clinical judgment, counselling skills, and organisation/efficiency in the headache-station and in the health-check-up station. In the intersection station and in the Diabetes station, they had problems in organisation/efficiency and medical interviewing skills. All results are reported in Table 1.

Bivariate correlation coefficients of all three instruments show strong correlations between mini-CEX and BGR and mini-CEX and checklists, respectively; and a slightly lower correlation between for between BGR and checklists (Table 2).

Differences in the OSCE scores between preclinical and clinical scores

Using Pillai's Trace in a MANOVA, we found a significant effect of the phases of studies on the scores of each of the measurement instruments in all four stations (Table 3). However, separate univariate ANOVAs on the outcome variables revealed non-significant effects for checklists on the intersection, $F(1,86) = 1.67$ and diabetes station, $F(1,86) = 0.43$, and for the BGR on the diabetes station, $F(1,86) = 2.97$, all $ps > 0.05$.

Students' evaluation of the GPr-OSCE

Overall, students were very pleased with the GPr-OSCE. All students found the patient problems relevant for clinical practice. The majority of our participants (90%) found the setting adequate to simulate a GPr consultation. Students' suggestions for improvements included the integration of a physical examination and a more authentic equipment of the rooms (e.g. computer, arrangement of chairs and tables). The majority of students would prefer a feedback of the SP directly after each encounter (85%) and a feedback of a GP after the whole OSCE (79%).

Discussion

Aim of the study was to pilot and test an OSCE specific for a General Practice setting. Based on previously conducted interviews with medical doctors, primarily GPs, and self-developed EPAs, we developed a GPr specific OSCE with four stations to test the effects of setting and stations with three different measurement instruments: mini-CEX, Berlin Global Rating (BGR) and self-developed content specific checklists. This is to our knowledge the first study, comparing two commonly used global ratings to content-specific checklists within an undergraduate student population in the field of General Practice.

Instead of a rotation of students through the different stations and rooms, we chose a novel more naturalistic setting, in which students stayed in one room and one SP after another entered their room. The purpose of this approach was to present the standardized problems as realistically as possible (Brannick et al. 2011) and allowed us to simulate a GP consultation hour. We calculated and compared the reliability of our three measurement instruments to find out if one of these could produce reliable results to use our OSCE in high-stakes exams. We hypothesized that clinical students would score higher than preclinical students in all measurement instruments within all stations as an expression for evidence of validity.

Although only four stations were included into our OSCE, inter-station reliability between stations was satisfactory using the BGR and mini-CEX. Reliability for using only our content specific checklists was low although items on checklists are usually rather easily observed. This might be due to the fact that our content-specific checklists were specifically designed to fit each station's circumstances rather than trying to measure an overarching construct. Downing et al. suggest that a reliability even for a low-stakes assessment ought to be above 0.70 (Downing 2004), which could only be achieved by the BGR and mini-CEX, but not by our checklists. Many studies assessing the reliability and validity of the mini-CEX were completed in Internal Medicine though recent studies have shown an implementation of the mini-CEX within multiple other disciplines but not yet in the area of General Practice (Humphrey-Murto et al. 2018).

As expected, clinical students scored better than preclinical students in all four stations using the mini-CEX, being an instrument used worldwide and tested for assessing overall clinical competence in undergraduate and graduate

Table 2. Bivariate correlation coefficients between OSCE scores.

	CL	Mini-CEX
BGR	0.515	0.755
CL		0.767

Table 1. Reliability and percentage of correct answers in the four OSCE-stations.

			Mean station score (%)			
	Cronbach's alpha	Mean overall test score (%)	Headache	Health-check-up	Intersection	Diabetes
BGR	0.88	62	66	64	61	59
CL	0.57	54	59	58	45	54
Mini-CEX	0.84	55	60	57	51	51

Table 3. MANOVA comparing phase of study on measurement instruments.

	Preclinic M (SD)	Clinic M (SD)	F	p
Pillai's trace approximate health check-up station $F(1,85) = 0.970$, $p < 0.05$				
BGR	9.12 (2.41)	10.95 (2.01)	14.306	<0.05
Checklists	16.59 (4.56)	19.73 (4.12)	10.819	<0.05
Mini-CEX	25.12 (8.95)	34.38 (8.13)	24.337	<0.05
Pillai's trace approximate headache station $F(1,86) = 0.971$, $p < 0.05$				
BGR	9.55 (2.71)	11.24 (2.12)	10.625	<0.05
Checklists	16.12 (4.32)	20.64 (3.82)	26.123	<0.05
Mini-CEX	25.76 (8.54)	36.24 (7.52)	36.162	<0.05
Pillai's trace approximate intersection station $F(1,86) = 0.958$, $p < 0.05$				
BGR	8.7 (2.58)	10.44 (2.51)	9.678	<0.05
Checklists	10.33 (3.05)	11.18 (2.94)	1.668	n.s.
Mini-CEX	21.91 (8.78)	30.84 (8.48)	22.254	<0.05
Pillai's trace approximate diabetes station $F(1,86) = 0.951$, $p < 0.05$				
BGR	8.82 (2.20)	9.75 (2.58)	2.971	n.s.
Checklists	11.67 (3.11)	12.18 (3.85)	0.424	n.s.
Mini-CEX	23.91 (9.02)	30.04 (8.33)	10.489	<0.05

scenarios. However, differences in scores of the content specific checklists were only significant for the 'headache-station' and the 'health check-up'. This might be due to the fact that these stations focus more on history taking and stating a diagnosis rather than sharing information and a counselling consultation. The inclusion of background information by the SP (diabetes diary, discharge letter) also demanded more flexibility and the ability to deal with complexity. History taking and differential diagnosis can both easily be trained in hospital settings although patient problems might be different from ambulant settings. Sharing information and counseling are more complex tasks. This makes them more difficult to be operationalized in checklists and more complex to train.

For the mini-CEX scores, all differences were significant between clinical and pre-clinical students. Even for the two stations, where content-specific checklists could not discriminate between clinical and pre-clinical students, mini-CEX produced significant results. While checklists are still the standard in OSCEs, our findings lead us to the conclusion that mini-CEX can be used for assessing medical students, also explaining more variance than our second global rating (BGR). Especially a field like GPr with many complex and ill-defined situations with uncertain outcomes and diagnoses, global rating like the mini-CEX seem to produce more sensitive results than checklists. These results also corroborate studies that have compared global ratings to checklists in other fields of medicine, showing global rating to be more efficient and reliable especially in high-stakes exams (Cohen et al. 1996; Regehr et al. 1999). Williams et al. (1999) also reported that the use of checklists can affect students' learning. Granularity of checklist items can result in superficial learning and recall of checklist items to receive credit for the station. This can lead to a rapid-fire questioning style and therefore counteract the purpose of the teaching scenario (Billings and Stoeckle 1999).

There are a number of limitations that should be considered when interpreting our findings. In contrary to most OSCEs, all stations in our study were scored by only one trained rater using video analysis. Van der Vleuten (1996) recommended the use of different assessors and observations to overcome problems with reliability in performance-based assessments. Further studies should include more raters to overcome this problem. Another way to contribute to more external validity could be the inclusion of SP ratings in our model. Cohen et al. (1996) already compared global SP ratings of interpersonal and communication skills to SP checklists.

Whereas our effects comparing checklists with global ratings were quite robust, one can still hypothesize that our checklists simply might be not complex enough to discriminate in complex situations. Even though most studies like ours reported good reliability for the mini-CEX, with internal consistency scores above .70, there are some studies noting lower values of just .44 (Cook et al. 2009). This could lead to a combined use of the mini-CEX with another instrument like the BGR for more robust results at least in high-stakes exams. Yet within our data the addition of one or two more instruments did not account for any additional variance when comparing clinical to pre-clinical students. While our findings point in a direction of excluding checklists, we could still see that the usage of content-specific checklists can be helpful in OSCE scenarios that are more focused on evaluating and developing technical skills (Morgan et al. 2001; Ringsted et al. 2003) or in early stages of education. However, the turning point at which level of medical training a switch from checklists to global ratings should be performed could not be answered with our data.

Feedback regarding our piloted OSCE showed that students took pleasure in a GPr specific scenario. For the inclusion in high-stakes exams, improvement of external validity should be the focus of further studies.

Conclusion

With the development of an EPA-based OSCE we tried to fill the gap of a missing practical assessment in GPr. When comparing behavioral checklists and global ratings within an undergraduate student population, we conclude to favor global ratings like the mini-CEX when testing clinical competence. This is in contradiction to a still common practice to rely on checklists in OSCE settings. To improve our OSCE, our piloted naturalistic setting should be maintained while taking measures to enhance external validity and reliability for high-stakes exams like formalizing extent and scope of our test.

Acknowledgements

We thank all medical doctors (namely Benita Mangold, Rolf Stegemann, Anne Simmenroth, Jürgen in der Schmitten), students,

simulated patients, and colleagues who helped us to conduct our study. We also thank our outstanding research assistants Katharina Schaefer, Katharina Blum, Claire Vogel, and Insa Wessels.

Ethical approval

The study protocol received approval by the Ethical Committee of the Medical Faculty of LMU Munich, Germany, proposal code 133-14.

Disclosure statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

Funding

The project was funded by the Dr. Hildegard-Hampp Foundation 2014.

Notes on contributors

Patrick Giemsa, MSc, is a Psychologist and Psychotherapist with his own practice. His area of research are social cognition and metacognition in patients with schizophrenia. He has also been lecturing statistics and research methods at MHB Neuruppin and SFU Berlin since 2014.

Clara Wübbolding, Dr. Med., studied medicine at the Ludwig-Maximilian-University until 2016. Central aspects of her dissertation related to different methods of assessing students especially in General Practice and Family Medicine. Since May 2017, she works as a resident physician at the 'Israelitisches Krankenhaus' in Hamburg.

Martin R. Fischer, MD, MME, FAMEE, is an internist and medical educator. He is a Professor of Medicine and the Director of the Institute for Medical Education at LMU Munich. He also serves as the Assoc. Dean for Clinical Studies at the Medical Faculty of LMU Munich.

Tanja Graupe (née Pander), Dipl. Päd., is research assistant at the Institute for Medical Education at the University Hospital, LMU Munich since 2012. Her main research topics are communication, emotion handling skills and empathy as well as scientific and clinical career paths and mentoring.

Anja Härtl, MD, is responsible for the faculty development at Department for Medical Education and Educational Research, Faculty of Medicine, University of Augsburg. Her main research interest is the development and integration of the role as an educator for health professions.

Christine Lenz, Dr. Med., is working as a General Practitioner in her own practice since 2001. She is specialized in Internal Medicine and General Practice. She lectures General Medicine at LMU Munich since 2004.

Linda Sanftenberg, Dr. rer. nat., is a research associate at the Institute of General Practice and Family Medicine, University Hospital, LMU, Munich, since 2014. The main topics of her studies are health promotion, preventive medicine, vaccination and immunization as well as travel medicine in General Practice.

Jörg Schelling, Dr. Med, MD, is working as a General Practitioner in a family-owned practice since 2006. He is specialized in Internal Medicine and General Practice and an Honorary Professor at LMU Munich. From 2014 to 2016 he was chair and founding director of the Institute for General Practice and Family Medicine of LMU.

Katrin Schüttpelz-Brauns, Dr. rer. nat., is the head of the Educational Research Team at the Department of Undergraduate Education and Educational Development, Medical Faculty Mannheim at Heidelberg University since 2012. Her main research topics are competency-based learning, teaching and assessing as well as formative assessments in medical education.

Claudia Kiessling, Dr. med., MPH, is a Professor in the Department of Health at the University of Witten/Herdecke and directs the curriculum for personal and professional development in medical education. Her research interests are clinical communication, patient-centred medicine, professional identity formation, and assessment.

References

Barrows HS. 1993. An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC. Acad Med. 68(6): 443–443.

Biggs J. 2003. Aligning teaching and assessing to course objectives. Conference on Teaching and Learning in Higher Education: New Trens and Innovations. University of Aveiro, Apr 13–17.

Billings J, Stoeckle J. 1999. The clinical encounter. A guide to the medical interview and case presentation. 2 ed. St. Louis: Mosby Press.

Brannick MT, Erol-Korkmaz HT, Prewett M. 2011. A systematic review of the reliability of objective structured clinical examination scores. Med Educ. 45(12):1181–1189.

Buddeberg-Fischer B, Stamm M, Buddeberg C, Klaghofer R. 2008. The new generation of family physicians-career motivation, life goals and work-life balance. Swiss Med Wkly. 138(21–22):305–312.

Buddeberg-Fischer B, Stamm M, Marty F. 2007. Family medicine in Switzerland: training experiences in medical school and residency. Fam Med. 39(9):651–655.

Cohen D, Colliver J, Marcy M, Fried E, Swartz M. 1996. Psychometric properties of a standardized patient checklist and rating-scale form used to assess interpersonal and communication skills. Acad Med. 71(1):89–97.

Cook D, Dupras D, Beckman T, Thomas K, Pankratz V. 2009. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. J Gen Intern Med. 24(1):74–79.

Cunnington J, Neville A, Norman G. 1996. the risks of thoroughness: reliability and validity of global ratings and checklists in an OSCE. Adv Health Sci Educ. 1(3):227–233.

DEGAM e.V. 2013. Musterlogbuch für das Praktische Jahr Allgemeinmedizin. Frankfurt am Main: DEGAM e.V. [accessed 2019 Jan 9]. https://www.degam.de/files/Inhalte/Degam-Inhalte/Sektionen_und_Arbeitsgruppen/Sektion_StudiumHochschule/Muster_PJ-Logbuch_3.1.pdf.

Downing SM. 2004. Reliability: on the reproducibility of assessment data. Med Educ. 38(9):1006–1012.

Duerson M, Romrell L, Stevens C. 2000. Impacting faculty teaching and student performance: nine years' experience with the Objective Structured Clinical Examination. Teach Learn Med. 12(4):176–182.

Gensichen J, Gerlach F. 2003. Allgemeinmedizinische Lehre – Herausforderungen der neuen Approbationsordnung. Zeitschrift Für Allgemeinmedizin. 79:405–407.

Gulich M. 2005. Praktisches Jahr in der Allgemeinmedizin – eine neue Herausforderung. Z Allg Med. 81(1):9–12.

Hasler L, Stamm M, Buddeberg-Fischer B. 2008. Future family physicians – reasons for their specialty choice and crucial professional skills. Praxis. 97(24):1277–1285.

Hodges B, Hanson M, McNaughton N, Regehr G. 2002. Creating, monitoring, and improving a psychiatry OSCE. Acad Psychiatry. 26(3): 134–161.

Humphrey-Murto S, Cote M, Pugh D, Wood TJ. 2018. Assessing the validity of a multidisciplinary mini-clinical evaluation exercise. Teach Learn Med. 30(2):152–161.

Ilgen J, Ma I, Hatala R, Cook D. 2015. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. Med Educ. 49(2):161–173.

Kopetsch T. 2010. Dem deutschen Gesundheitswesen gehen die Ärzte aus! Studie zur Altersstruktur- und Arztzahlentwicklung. 5 ed. Berlin: Bundesärztekammer und Kassenärztliche Bundesvereinigung.

Morgan P, Cleave-Hogg D, Guest C. 2001. A comparison of global ratings and checklist scores from an undergraduate assessment using an anesthesia simulator. Acad Med. 76(10):1053–1055.

Norcini JJ, Blank LL, Duffy FD, Fortna GS. 2003. The Mini-CEX: a method for assessing clinical skills. Ann Intern Med. 138(6):476–481.

Norcini J, Burch V. 2007. Workplace-based assessment as an educational tool: AMEE Guide No. 31. Med Teach. 29(9–10):855–871.

Peters H, Holzhausen Y, Boscardin C, Ten Cate O, Chen H. 2017. Twelve tips for the implementation of EPAs for assessment and entrustment decisions. Med Teach. 39(8):802–807.

Regehr G, Freeman R, Robb A, Missiha N, Heisey R. 1999. OSCE performance evaluations made by standardized patients: comparing checklist and global rating scores. Acad Med. 74(10):135–137.

Ringsted C, Ostergaard D, Ravn L, Pedersen J, Berlac P, van der Vleuten C. 2003. A feasibility study comparing checklists and global rating forms to assess resident performance in clinical skills. Med Teach. 25(6):654–658.

Scheffer S, Muehlinghaus I, Froehmel A, Ortwein H. 2008. Assessing students' communication skills: validation of a global rating. Adv Health Sci Educ. 13(5):583–592.

Swayamprakasam A, Segaran A, Allery L. 2014. Work-based assessments: making the transition from participation to engagement. JRSM Open. 5(3):204253331351586.

Ten Cate O. 2017. Entrustment decisions: bringing the patient into the assessment equation. Acad Med. 92(6):736–738.

Ten Cate O, Graafmans L, Posthumus I, Welink L, van Dijk M. 2018. The EPA-based Utrecht undergraduate clinical curriculum: development and implementation. Med Teach. 40(5):506–513.

Townsend A, McIlvenny S, Miller C, Dunn E. 2001. The use of an objective structured clinical examination (OSCE) for formative and summative assessment in a general practice clinical attachment and its relationship to final medical school examination performance. Med Educ. 35(9):841–846.

Turner J, Dankoski M. 2008. Objective structured clinical exams: a critical review. Fam Med. 40(8):574–578.

Van der Vleuten C. 1996. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ Theory Pract. 1(1):41–67.

Walker H. 2006. Primary care is dying in the United States: mutatis mutandis. Med Educ. 40(1):9–11.

Whitcomb M, Cohen J. 2004. The future of primary care medicine. N Engl J Med. 351(7):710–712.

Wilkinson T, Frampton C, Thompson-Fawcett M, Egan T. 2003. Objectivity in objective structured clinical examinations: checklists are no substitute for examiner commitment. Acad Med. 78(2):219–223.

Williams R, McLaughlin M, Eulenberg B, Hurm M, Nendaz M. 1999. The patient findings questionnaire: one solution to an important standardized patient examination problem. Acad Med. 74(10):1118–1124.