Assessment methods in medical specialist assessments in the DACH region – overview, critical examination and recommendations for further development

Nils Thiessen[1]
Martin R. Fischer[2]
Sören Huwendiek[3]

1 EDU - a degree smarter, Digital Education Holdings Ltd., Kalkara, Republic of Malta

2 LMU München, Klinikum der Universität München, Institut für Didaktik und Ausbildungsforschung in der Medizin, München, Germany

3 Universität Bern, Institut für Medizinische Lehre, Abteilung für Assessment und Evaluation, Bern, Switzerland

Abstract

Introduction: Specialist medical assessments fulfil the task of ensuring that physicians have the clinical competence to independently represent their field and provide the best possible care to patients, taking into account the current state of knowledge. To date, there are no comprehensive reports on the status of specialist assessments in the German-speaking countries (DACH). For that reason, the assessment methods used in the DACH region are compiled and critically evaluated in this article, and recommendations for further development are described.
Methods: The websites of the following institutions were searched for information regarding testing methods used and the organisation of specialist examinations:

1. Homepage of the Swiss Institute for Medical Continuing Education (SIWF),
2. Homepage of the Academy of Physicians (Austria) and
3. Homepage of the German Federal Medical Association (BAEK).

Further links were considered and the results were presented in tabular form. The assessment methods used in the specialist assessments are critically examined with regard to established quality criteria and recommendations for the further development of the specialist assessments are derived from these.
Results: The following assessment methods are already used in Switzerland and Austria: written examinations with multiple choice and short answer questions, structured oral examinations, the Script Concordance Test (SCT) and the Objective Structured Clinical Examination (OSCE). In some cases, these assessment methods are combined (triangulation). In Germany, on the other hand, the oral examination has so far been conducted in an unstructured manner in the form of a 'collegial content discussion'. In order to test knowledge, practical and communicative competences equally, it is recommended to implement a triangulation of methods and follow the further recommendations described in this article.
Conclusion: While there are already accepted approaches for quality-assured and competence-based specialist assessments in Switzerland and Austria at present, there is still a long way to go in Germany. Following the recommendations presented in this article, a contribution could be made to improving the specialist assessments in the DACH region according to the specialist assessments objectives.

Keywords: medical specialist assessment, DACH region, cognitive, practical and communicative competencies

Introduction

Assessments fulfil a wide range of important tasks: they have a strong effect on learning, they can be used to provide feedback on the effectiveness of education and training programmes and, consequently, they can help protect patients [1]. Until the middle of the twentieth century, medical assessments were mainly written in the form of essays or oral assessments [1], [2]. At that time, evaluations derived from assessments often turned out to be subjective, arbitrary and non-reproducible [1]. Subsequently, standardised tests such as multiple-choice examinations (MC examinations) or Objective Structured Clinical Examinations (OSCE) [3] were developed (Case & Swanson 1996, Norcini & Burch 2007, Kogan et al. 2009, cited after Norcini [1]). Tests should be objective, reproducible (reliable) and valid. Furthermore, they should be accepted by test takers and examiners, have a learning-promoting component and be as cost-efficient as possible [4]. "Objectivity" means that the test should be as independent as possible from the examiner as a person – their attitudes, feelings and motives. It refers to the performance, evaluation and interpretation of a test [5]. A test should provide approximately the same result when repeated; in other words it should be "reliable". Reliability is a measure of the trustworthiness of a test. Reliability is represented as a coefficient ranging from 0 (no reliability) to 1 (perfect reliability). The value 0.80 is often set as the minimum standard for a significant high stakes test [6]. Among other things, a test is "valid" if it measures what it claims to measure. It is thus a measure of the measurement accuracy of a test [5]. It would be desirable for all valid tests to be compared to an external standard, but one is often not available in practice. In this case, expert assessments are often used for validation. In medical education, constructs are mainly used – in other words, abstract concepts and principles derived from behaviour and explained by pedagogical and psychological theories [7]. This fact is represented by the concept of "construct validity".

Society relies on tests that ensure that patients can place themselves in the care of competent and qualified physicians who have reached a minimum standard [1]. According to Premi, specialist examinations should ensure that colleagues who have passed this assessment have acquired the knowledge and necessary skills of their specialist group and can apply them independently (Premi 1994, quoted from Ratnapalan & Hilliard [8]. Examinations that demonstrate the necessary knowledge, skills and attitudes for the pursuit of a profession are part of self-regulation in continuing medical education. This is viewed with increasing scepticism worldwide, especially against the background that training to become a physician is a very expensive affair and is often financed by the public purse [9]. For this reason, governments in Australia, Great Britain and Canada are directly entrusted with regulating continuing medical education (Chantler & Ashton 2009, Shaw et al. 2009, Medicare Advisory Commission 2009, cited after Holmboe [9]).

The official training regulations for postgraduate medical doctors, which were developed by the BAEK, have the character of a recommendation. Completion of the specialist training is assessed on the basis of documented competences, issued by the respective physician in charge, and an oral examination. According to the BAEK, this "certificate of further training" is proof of the acquired competence and serves as quality assurance for patient care and citizen orientation. The term 'competence' is not specified here [10].

In Switzerland, the FMH (Foederatio Medicorum Helveticorum) is the professional association of Swiss physicians. It is the umbrella organisation of more than 70 doctors' organisations. The Swiss Institute for Medical Continuing Education (SIWF) is an autonomous body of the FMH and ensures high-quality continuing education for doctors in over 120 specialist areas. In cooperation with the professional associations, the SIWF issues a detailed further training programme [https://www.siwf.ch/] for each specialist area.

The Austrian Medical Association (ÖÄK) grants the right to practise as employed, self-employed and self-reliant physicians. The ÖÄK has entrusted the 'Austrian Academy for Physicians GmbH [https://www.aerztekammer.at/] with the implementation of the medical examination as a prerequisite for the pursuit of a medical profession. The training contents and the corresponding certificates for the acquisition of a specialist title have been drawn up and specified by the ÖÄK [https://www.aerztekammer.at/ausbildungsinhalte-und-rasterzeugnisse-kef-und-rz-v-2015]. Over the last few years, competence orientation has increasingly come to the fore in medical education and postgraduate medical education. This is with the aim of ensuring that graduates master the challenges of practical work and possess all the necessary skills [11]. The physician competency framework "Canadian Medical Education Directives for Specialists" (CanMEDS) was developed by Frank et al. to guarantee comprehensive postgraduate training for physician [12]. On the basis of a systematic literature analysis and broad-based expert and stakeholder surveys, seven medical roles were defined by Frank et al. [11] to establish CanMEDS and integrate it into all of Canada's continuing education programmes [13]:

1. Medical Expert
2. Communicator
3. Collaborator
4. Leader
5. Health Advocate
6. Scholar
7. Professional

Numerous key competencies are assigned to these roles. The CanMEDS role model has already been integrated into national learning objective catalogues for medical studies in Europe (Netherlands (Laan 2010, quoted from Jilg) [13], Switzerland [14], Germany [http://www.nklm.de]). Not only should medical teaching be competence-oriented, but also any successive further

education. This includes a competence-oriented examination of knowledge, skills and attitudes.
To date, there exists no compilation of the extent to which the specialist assessments of the DACH region are competence-based and whether the quality criteria previously mentioned, such as objectivity and reliability, are taken into account. The aim of this work is therefore to provide an overview of the existing summative specialist assessment formats in the DACH region and their organisation, to take a critical look at the formats with regard to quality criteria and to make recommendations on the basis of the international literature. As a first step, this compilation should make the current situation better known and highlight possible directions for the further development of specialist assessments in the DACH region.

Methods

The following homepages were searched for references to existing assessment formats and the organisation of specialist assessments:

1. Homepage of the Swiss Institute for Medical Education and Training [https://www.siwf.ch/]
2. Homepage of the Austrian Academy of Physicians [https://www.aerztekammer.at/]
3. Homepage of the German Federal Medical Association [10]

Using the websites of these national umbrella organisations, the contents of the homepages of professional associations or regional chambers of physicians were evaluated. These provided further information on the examination formats currently used. The quality of the Internet research therefore depends on the information and data listed there. A review of the individual annual reports of the State Medical Associations provided, as far as available, an overview of the number of examinations carried out in one year and the corresponding failure rates. Further key statistical figures or costs were not provided.
Furthermore, the test methods used were critically evaluated with regard to the quality criteria of the tests (validity, reliability, objectivity, acceptance, cost efficiency and influence on learning). In addition, criteria for best practice specialist assessments were derived from the literature.

Results

Overview of the assessment formats used in the DACH region

Switzerland

In Switzerland, Multiple Choice (MC) examinations are used in 26 of 46 subject areas. The number of questions varies between 50 and 200. The minimum exam duration is 120 minutes; the maximum is 360 minutes. The specialist areas of Anaesthesiology, Allergology/Immunology, Cardiology and Vascular Surgery are examined in conjunction with the European Union of Medical Specialists (UEMS). The following types of questions are used in Switzerland, although there is a clear variance in their occurrence within the specialist areas: type A positive, type A negative, type Kprim, type B, type E, type R, and type Pick N are used as MC formats. Short Answer Questions (SAQ) and Script Concordance Tests (SCT) are also used. Swanson and Case [15] provide a good introduction to this, including examples for the different question types. In the field of Psychiatry/Psychotherapy, candidates must submit a written paper in part 2 and take part in a colloquium. In Radiology, for example, open text formats are used, but they are not explained in detail. Here the diagnosis of cases is in the foreground and a web-based examination tool is used. In addition to the written format, there are also oral assessments in Switzerland, including discussion of a paper, presentation of a patient case, holding a colloquium and structured oral assessments (SMP). The duration of these varies from 20 minutes to 180 minutes. Some subjects, e.g. Endocrinology/Diabetology, combine a written examination with an oral examination. In 23 subject areas, i.e. 50% of the specialist assessments in Switzerland, assessments with a practical component take place. In the fields of Oto-Rhino-Laryngology and Thoracic Surgery, for example, practical examinations are held as part of an operation. Rheumatology carries out an OSCE comprising 9 stations, with 10 minutes available per station. It is not only knowledge (Anatomy, Pathophysiology etc.), but also practical skills (examination techniques), as well as communicative skills, that are tested in a standardised way. In 2017, the SIWF awarded 1428 medical specialist titles [16] and, in 2018, 1434 medical specialist titles [17]. The SIWF does not publish a failure rate in its annual report.

Austria

14 of 57 tests are performed with MC questions. The minimum number of questions is 50 and they can reach up to 200 questions, which have to be answered in the field of skin and venereal diseases. The candidates have 60 to 300 minutes at their disposal. Anaesthesiology and Urology are affiliated to the European specialist examination. The Pathology and Radiology departments use tests with short answer questions, which last from 80 to 240 minutes. In Austria, 45 subject areas are examined orally on the basis of structured oral examinations (the use of a so-called “blueprint”: pre-formulated questions and a horizon of expectation). In this context, the term “blueprint” refers to a weighted assessment plan in which the selection of relevant examination content ensures that each candidate is treated equally in terms of that content. For most subjects, a blueprint is created and explicitly mentioned in the exam description. The duration of the examination can vary between 40 and 120 minutes. Some subjects are examined both in writing and orally. There are currently no clinical-practical examinations in

Austria. We do not have statistics on the number of tests carried out in Austria per year and the corresponding failure rates.

Germany

The specialist assessment is held at all regional medical associations in the form of an unstructured oral examination (UMP), which lasts at least 30 minutes and can last up to 60 minutes. This type of examination is used for all medical specialist qualifications and is also referred to as a "collegial expert discussion". The number of examiners may vary and at least one examiner must be from the field to be examined. The examination results must be documented. Typically, a structured blueprint is not prepared and the questions are not pre-formulated in advance (in the sense of a standardised and structured examination and a specified expectation horizon). The Landesärztekammer Hamburg is an exception: here, the questions are handed over in advance to the chairman of the examination. Table 1 shows the number of specialist examinations and the associated failure rates. The data were taken from the annual reports of the respective regional medical associations for the year 2017, which were available online. An inquiry to the BAEK regarding comprehensive, nationwide statistics revealed that such statistics were not available.

Critical appraisal of the tests used in the DACH region

MC exams

MC examinations are widely used in medicine as an assessment method because they can be cost-efficient and can offer high validity and reliability for testing knowledge (Norcini 1985, cited after Gerhard-Szep [18]. This presupposes, however, that a sufficient number (at least 40) of high-quality questions (in content and form) are used per test (Jünger 2014, cited after Gerhard-Szep [18]). Case et al. emphasise that two criteria are necessary to develop a good question: the question must both examine relevant content and be well structured [15]. The development of MC questions at a qualitatively high level is time-consuming. With written examination methods it is above all possible to test factual knowledge. In contrast to an OSCE, it is not possible to test communicative and practical skills, or competences, using MC questions [19].

Short answer questions

For short answer questions, freely formulated, short, keyword-like answers must be given. Test takers must spontaneously think of the correct solution and cannot react to given answers [5]. This reduces the so-called "cueing" that gives candidates the opportunity to answer a question correctly without knowledge (Schuwirth 2004, cited from Epstein [20]). Ideally, "context-rich" question strains (case vignettes) are also offered here, which make it possible to test application knowledge and, for example, "clinical reasoning". Reliability also depends to a large extent on the quality of evaluations carried out by the examiners [20] – in this case, training the examiners in advance can help. The evaluation is more susceptible to subjective distortions than with MC questions. Pre-formulated expectation horizons, to which the evaluators must orient themselves, can increase objectivity and should be available. Acceptable reliability values can be achieved by using several testers, each of whom is responsible for evaluating different tasks [5]. Rademakers et al. [21] provide a clear presentation of a task. In the meantime, there is also the possibility to evaluate computer-based answer options [22]. In the near future, new developments are to be expected in this area, which will make use of artificial intelligence methods.

Script Concordance Test (SCT)

An SCT is used to check the 'clinical reasoning' competence of examinees in situations of clinical uncertainty [23]. Short clinical scenarios are described and additional information is provided step by step. In light of this new information, the investigator should then make diagnostic, follow-up or therapeutic decisions [24]. Using a 5-point Likert scale from -2 to +2, the examinee must indicate to what extent the additional information supports or does not support the disease hypothesis described in the scenario [25]. The results of the test takers are subsequently compared with the assessments of an expert group; the "gold standard" answer achieves the greatest number of points on which most experts have agreed [23]. Figure 1 shows an example of three questions [26]. Various working groups have been able to demonstrate the favourable psychometric properties of the SCT (construct validity, reliability and feasibility) [24]. Brailovsky et al. (2001, quoted after Epstein [20]) were able to show that the answers to such questions correlate with the candidate's level of education and can predict their future performance in oral examinations in terms of their "clinical reasoning" ability [20]. A critical weakness of the 5-point Likert scale is that it can lead to misunderstandings and false assessments by the expert panel, so Lineberry et al. recommend the use of a 3-point scale consisting of the following: "refuted", "neither refuted nor supported" and "applicable". In addition, there is a risk that candidates' answers will tend towards the middle and thus they will obtain a better test result than those who use the Likert scale in its extremes [26]. In addition, the usefulness of scores corresponding to an expert group is still under discussion, especially since 10-20 members [27] are recommended for this. The SCT is therefore quite complex.

Structured Oral Examination

The oral examination is a traditional form of examination in which one or more examiners address questions to the candidate. The oral exam is designed to evaluate know-

Table 1: Overview of the key audit figures in the annual reports of the regional medical associations for the year 2017.

	Assessments performed	Passed assessments	Failure rate in %
Medical Council Berlin	1256	1182	6
Medical Council North Rhine	1731	1651	4,6
Medical Council Hamburg	513	501	2,3
Medical Council Brandenburg	322	299	7,1
Medical Council Bremen	127	122	3,9
Medical Council Saarland	185	179	3,2
Medical Council Saxony-Anhalt	314	298	5
Medical Council Westphalia-Lippe	1288	1182	8,2
			5,0375

You are thinking of a	You then receive the following additional information:	Your hypothesis is now*				
1. Cerebral abscess	Patient had middle ear infection 10 days ago.	-2	-1	0	+1	+2
2. Ischemic stroke	Sudden onset of symptoms 2 hours ago	-2	-1	0	+1	+2
3. Cerebral metastasis	Normal CCT with contrast medium	-2	-1	0	+1	+2

*-2 = incorrect, -1 = unlikely, 0 = neither more probable, nor less probable, +1 = more probable, +2 = certainly true
CCT = Cranial computer tomogram

Figure 1: Case 1: A 75-year-old man presents with a right hemiparesis in the emergency room.

ledge, explore depth of knowledge and test other qualities such as mental agility. Colton & Peterson, Foster et al. and Kelly et al. criticised the use of oral examinations in high-stakes tests because of their low reliability (Peterson 1967, Foster et al. 1969, Kelly et al. 1971, cited after Davis [28]). During an oral examination numerous sources of error occur to which examiners are subject in the framework. For example, with the primacy effect, first impressions dominate over later impressions and, with the recency effect, later impressions are more lasting. In the halo effect, the perception and evaluation of one property outshines the perception and evaluation of other properties. Antipathy, sympathy and the composition of the examiners also have an influence on the evaluation of the test performance [29]. According to Roloff et al. reliability and objectivity increase when several examiners examine independently and the number of questions and the examination time increase (Roloff 2016, cited by Gerhard-Szep [18]). Memon et al. have named 15 quality assurance measures that are necessary from the point of view of the literature to ensure the objectivity, reliability and validity of specialist examinations. They state that the oral examination is most suitable for clinical reasoning and decision making. The content of the examination should be determined in advance by a panel of experts. The examination questions should be selected in such a way that they adequately examine not only the corresponding depth of knowledge but also the breadth of the subject area and guarantee a corresponding inter-item reliability. Examiners must first be trained with regard to carrying out an oral examination. Deviations between examiners (inter-examiner variations) must be monitored and addressed. Item creation and implementation processes must be standardised and a statistical evaluation should give conclusions about reliability. In the case of oral examinations, bias must be expected in the assessment and therefore quality assurance should be carried out to this end [30].

Unstructured Oral Examination

An unstructured oral examination is usually carried out by two untrained examiners who examine based on their experience. Typically, there is neither a pre-formulated expectation horizon nor previously written questions based on the curriculum or blueprint. As early as 1985, Jayawickramarajah et al. were able to demonstrate that two thirds of the questions in an unstructured oral examination exclusively examined factual knowledge. An additional problem of an unstructured oral examination is the high probability of an occurrence of Construct Irrelevant Variance (CIV) due to the fact that too few examiners are used. CIV could occur, for example, when the testing of the competence “clinical decision making” is influenced by the appearance, fear, language skills or clothing of the examinee. Construct Underrepresentation (CU) is a further hurdle that must be considered in the context of an unstructured oral examination, since, for example, two to three clinical scenarios that are tested cannot cover the entire range of the substance area to be tested. Concerns about the validity of this traditional form of examination have led to it being replaced by written examinations or structured oral examinations (Jayawickramarajah et al. 1985, Turnball, Danoff, & Norman, 1996, Pokorny & Frazier, 1966, quoted from Lamping et al. 2007 [31]).

OSCE

The OSCE test format was developed by Harden in the 1970s and primarily tests clinical and practical competencies. A higher objectivity is achieved through standardisation [3]. In order to achieve a standardised presentation of illnesses, actors are specially trained for this purpose [32]. A number of problems are presented to the examinee in the form of a course. Both the number of stations and the number of examiners have a positive effect on its reliability. Despite high variance in studies, a good reliability can already be achieved with more than 10 stations [33]. The examinee has approximately 5-15 minutes per station to complete the task [2]. The investigator checks the observed clinical competence using a checklist and/or a global rating scale. OSCEs allow a diagnosis to be made in the context of contact with the standardised patient (SP) through skilful anamnesis techniques and a patient-centred physical examination. According to Van der Vleuten and Tamblyn, trained SPs cannot be distinguished from real patients; they can repeatedly perform reliably, and they can also give valuable feedback to the test subjects (Van der Vleuten 1990, Tamblyn 1991, cited after Newble [2]). The OSCE is generally regarded positively by students and lecturers (Roberts & Brown 1990, quoted from Rushfort [34]), even if some students perceive the exam as stressful. Compared to other exams, it is more objective (Schuwirth & Van der Vleuten 2003, quoted after Rushforth [34]) and has a positive effect on the motivation of the candidates to study for it (Bartfey et al. 2004, quoted after Rushforth [34]).

Overview of the evaluation of the examination formats

Table 2, following Gerhard-Szep et al., Lubarsky et al., Epstein et al. and van der Vleuten et al. [4], [18], [20], [25] below, was designed to illustrate the most frequently occurring forms of assessment in the DACH region and to classify them from the authors' point of view with regard to the essential quality criteria described. It essentially serves to provide a better overview and is intended to support the recommendations for a best practice specialist assessment.

Recommendation of a Best Practice Specialist Assessment

In the following, recommendations for a best practice specialist assessment are given. These have been derived from the current literature.

Observance of the following recommendations helps to ensure that the resulting tests are as valid, reliable, objective, accepted, instructive and cost-effective as possible. This is necessary so that competence-oriented learning objectives can be meaningfully tested and examinations can prove that candidates have learnt the competences necessary for the independent treatment of patients.

1. **Use of different test methods (triangulation):** Different test methods should be used to adequately test knowledge, on the one hand, and practical and communicative skills on the other. Only the combination (triangulation) of the results of different assessment formats can ensure high validity and different competences [35].
2. **Prior definition of the contents and competences to be tested (blueprinting):** A weighted examination plan (so-called blueprint) provides a framework for the assessment by ensuring that a balanced selection of relevant learning objectives is incorporated into the test before it is held [36]. This is to ensure that the test is valid, fair, relevant and representative of the subject being examined.
3. **Prior definition of the questions and the horizon of expectations:** For oral and practical examinations, as for written formats, the questions and the horizon of expectations must be recorded in writing for each question/station in advance (so-called structuring). In oral and practical assessments, clearly structured checklists unequivocally present the horizon of expectations and thus ensure the necessary objectivity of interpretation and evaluation [18], [37].
4. **Sufficient number of questions, examiners, stations/learning objectives to be examined:** A minimum reliability of 0.8 is given for relevant assessments [4], [35]. In order to improve these, the number of tasks and/or their quality can be increased [35]. Likewise, the number of examiners has a positive effect on reliability (Swanson 1987, cited after Lynch) [38]. The more examiners test, the better the reliability becomes. In oral and practical examinations, it makes more sense to have one examiner per topic/station rather than several examiners at the same time with fewer stations/subjects.
5. **Quality assurance of the created questions/tasks:** The content and formal linguistic review and revision of the tasks is necessary to guarantee the unambiguity of the answers and the high quality of the tasks and questions. The validity of the examination results is strengthened by a review process, where experts trained in medical didactics review the questions and tasks [4], [39].
6. **Quality assurance in the evaluation of the assessment:** Quality assurance through test statistical evaluation of examinations makes it possible to revise OSCE stations and examination tasks in a targeted manner, to examine checklists and, if necessary, to draw conclusions about the quality of teaching. The following parameters are recommended: for written assessments at least one evaluation should be carried out with regard to reliability, selectivity and item difficulty (except for small numbers of candidates – i.e. less than 30 – because of the influence of chance). For oral or practical examinations, the better "OSCE-

Table 2: Presentation of the assessment methods used in the DACH region with regard to relevant characteristics, from "++" (=high) to "—" (=low) or suitable to unsuitable from the authors' point of view.

	Objectivity	Reliability *	Outlay	Knowledge testing	Practical skills assessment
MC - Examinations	++	++	+	++	--
SMP	++	++	++	++	--
OSCE	++	++	++	+	++
SCT	++	++	+	++	--
SAQ	++	++	+	++	--
UMP	--	--	--	-	--

* = depending on number and quality of examination tasks

Metrics" [40] are very desirable for the evaluation of selectivity and item difficulty, as well as reliability at item level. A more modern approach to determining the pass mark for examinations, in which passing can also take place with more or less than 50% solved tasks, is that where the pass mark is determined in terms of content [36] – for example, a modified Angoff procedure with MC [41], or the borderline regression method with OSCE (see Wood et al. [42]).

7. **Learning effect for the candidates:** Assessments do not only serve for decision-making but they are also very important as a learning incentive for candidates and, additionally, they support the learning effect by giving the candidates feedback regarding their examination results [4]. For example, a feedback letter can be designed in such a way that the candidates know which areas of the blueprint they did less well in compared to the other tasks and the other candidates.

Consideration of cost efficiency

High quality assessments have their price, but they definitely represent a worthwhile investment with regard to the learning effect of test items [4], [39]. The method of examination should be chosen in each case, based on both its ability to adequately examine the subject matter (content validity) and it being as cost-efficient as possible. If, for example, it is primarily a question of testing the application knowledge of many candidates, a written examination with vignette questions is superior to a structured oral examination in terms of cost efficiency. The merger of professional societies can reduce the effort involved in, for example, practical examinations (like OSCE), with the aim of checking the CanMEDS roles (cf. the Swiss basic examination in surgery in the field of knowledge https://basisexamen.ch/).

Discussion

In this paper, the question of which specialist assessment methods are used in the DACH region is examined. In addition, the assessment methods used are critically reviewed and recommendations for the further development of specialist assessments are described, based on the current literature.

Testing methods used

More than 50% of the specialist assessments conducted in Switzerland take the form of MC examinations. The specialist areas of Anaesthesiology, Allergology/Immunology, Cardiology and Vascular Surgery are examined in conjunction with the European Union of Medical Specialists (UEMS). Other departments are planning to do so. Seven different types of MC questions are used, as well as the SAQ, free text examinations (not specified in more detail) and the SQT. Written work, SMP, practical examinations and an OSCE are also used. In total, 50% of the Swiss specialist examinations have a practical component (cf. attachment 1 and attachment 2).
In Austria, 25% of the specialist assessments are conducted as written examinations (MC questions). Two specialist areas examine in conjunction with the UEMS. SAQ and SMP are also related forms of examination. Blueprinting is used regularly. There is no practical examination yet (cf. attachment 2 and attachment 3). In Germany, an unstructured oral assessment takes place throughout the country, which is referred to as a "collegial content discussion" (cf. attachment 4).

Critical consideration and recommendations for the further development of specialist assessments

It is positive to note that in Switzerland 50% of the specialist assessments already have a practical component.

In order to be able to test practical and communicative competences within the scope of the specialist assessment, a practical, communication and competence-oriented examination should be used in addition to a knowledge-oriented examination method (written/SMP). The OSCE format could be used here. In order to reduce costs, for example, at least parts of the examinations could be conducted nationwide. According to the literature, it is also positive to note that, in Switzerland and Austria, MC tests with type A pos. questions are used in the majority of cases to objectively test knowledge – including application knowledge – and evaluate it statistically.

However, an assessment of the examined competence using these forms of examination is only possible if a detailed insight into the examinations is conducted and their results can be provided. The preparation of written examinations is often underestimated, and it is time-consuming, as review processes must take place in terms of content, formal language and (medical) didactics in order to guarantee the unambiguity of the answers provided. The workload is mainly shifted to the preparation phase (cf. Gerhard-Szep) [18]. In addition, ethical and cultural questions are often avoided when questions are created, since context-rich questions are difficult to write (Frederiksen 1984, cited from Epstein) [20]. Swing et al. generally recommend the use of regular examiner training as well as the use of expert groups that regularly critically question the examination method used [43]. Application knowledge does not have to be examined in writing, but it can also be examined in a structured, oral way. In addition to application knowledge, structured oral assessments can evaluate clinical decision making, professional thinking and self-confidence [30]. It should be borne in mind that every structured oral assessment is associated with high costs due to the high space and personnel requirements. Blueprinting and the creation of an expectation horizon are also necessary. Duration, number and experience of the examiners have a direct influence on the quality criteria of the structured oral assessment (Roloff 2016, cited after Gerhard-Szep [18]). Examiner training can help to raise awareness and reduce the psychological sources of error (see Kugler 2007 [29]) to which examiners are unconsciously subject. Here, resistance must be expected from previous examiners, who have tested for years without having been trained to do so.

In Germany, oral examinations are conducted in an unstructured manner, although the German Medical Association expressly emphasises on its website that the continuing education designation serves as proof of the acquired competence and quality assurance of patient care and citizen orientation [44]. The current specialist examinations in Germany therefore do not meet this requirement, as unstructured oral assessments do not fulfil the required quality criteria (also see table 2). It can therefore only be assumed that the UMP has been developed following a tradition and has not yet been subject to a critical review. Therefore, the unstructured oral assessments cannot be recommended. In order to better fulfil their responsibility for medical quality assurance in the future, a first step could be to establish contact with medical faculties that have been gathering experience in the implementation of structured examination methods for several years. An exchange of experience could also take place with experts from the DACH region who are already carrying out quality assurance for the specialist assessments used. In order to convert unstructured oral examinations into structured oral examinations at short notice, the regional medical associations in Germany could get in touch with colleagues trained in medical didactics to conduct examiner training on site. Flum also emphasises the aspect of quality assurance by saying that it would be helpful to standardise competences and testing methods in postgraduate medical training in general medicine within the EU in order to ensure the quality of treatment and patient safety (45). Likewise, specialist assessments should be seen as an instrument for regulating the content of continuing education and should meet the actual care needs of the population as well as the learning objectives and curriculum [45].

A recommendation with regard to a best practice specialist assessment is made against the background that, so far, only a few articles have taken a stand on assessment methods in the field of postgraduate medical education. A combined use of assessment methods (triangulation) is indispensable in order to be able to cover the necessary competence spectrum. Likewise, the following should be used and documented: obligatory blueprinting (Dauphinee 1994, cited from Wass), the preparation of questions and horizon of expectations in advance, a sufficient number of questions/tasks, quality assurance measures relating to the preparation and evaluation of examination questions/tasks, examination feedback (Gronlund 1998, cited from Norcini) and the most cost-effective assessment methods possible. These recommendations are supported by numerous publications [1], [37], [46]. Caraccio et al. emphasise that different assessment methods should be combined so that the competence level of continuing training assistants can be assessed (Caraccio 2013, quoted from Flum [45]. Taylor et al. point to a necessary standardisation of specialist assessments, the costs of which must, however, be considered (Taylor 1998, cited after Flum [46]). David et al. and Adler et al. rightly discuss at this point that costs incurred in the context of continuing training must be reflected in the Diagnosis Related Group (DRG) system [47], [48]. The necessary use of blueprinting in the field of postgraduate medical education is supported by Wass et al. In order to optimise learning success during postgraduate medical education, competences should be continuously recorded using various methods and feedback given through formative assessments (e.g. Mini Clinical Examination (Mini-CEX), Direct Observation of Procedural Skills (DOPS), Portfolios etc.) [1], [49]. Competency-based curricula should take into account how knowledge, skills and attitudes are tested at the highest examination level: “does”, according to Miller (Miller 1990

[50]). Examinations should be competence-based [11]. Currently, so-called EPAs (Entrustable Professional Activities) are increasingly being used to support curriculum development and test new ways of competence-based learning and testing [51]. Such a professional activity (EPA) could be, for example, the identification of an emergency patient on a normal ward and the initial assessment and initiation of necessary medical measures. A special feature of the EPA approach is the assessment of the learner on the basis of the presumed need for supervision ("entrustment"). The use of EPAs is often intuitively attractive for clinically active physicians, but their potential, including existing challenges (see also literature on workplace-based assessments [51], must be further investigated before the replacement of quality-assured summative specialist examinations could be considered.

On the one hand, this work is certainly limited by the fact that the data researched on the Internet can only be presented descriptively. On the other hand, continuing education curricula, the implementation of continuing education and the design of examinations are closely interlinked in the sense of "constructive alignment". A further limitation is therefore that, in this article, we have primarily dealt with the presentation of the examinations used in the DACH region and that curricula on further specialist training could only be mentioned in passing. However, a follow-up article could explain the different curricula on further specialist training and their implementation within the DACH region.

Conclusions

In the DACH region, the organisation of specialist assessments, the assessment methods used and the quality assurance measures are very different. In contrast to Germany, structured and standardised specialist assessment methods are already used in Austria and Switzerland – as well as practical examinations in the latter. If specialist assessments are to ensure that specialist doctors have the necessary competences for patient care, they must also be designed in such a way that they can actually test competences. This currently appears not to be the case in all specialist areas in Germany, but also in most specialist areas in Austria and Switzerland. Therefore, in order to ensure the quality of postgraduate medical training, it is necessary that even more attention is paid in the three countries to the fact that summative specialist assessments also examine the intended competences of the prospective specialists. A combination of a written examination with a practical examination (e.g. OSCE) is currently recommended, as this not only tests knowledge but also other competences including practical and communicative skills (see table 2).

Acknowledgements

We thank Dr. med. Susanne Frankenhauser, MME, Dr. med. Uta Krämer and Brian Webber for their constructive help in reviewing the manuscript.

Competing interests

The authors declare that they have no competing interests.

Attachments

Available from
https://www.egms.de/en/journals/zma/2019-36/zma001286.shtml

1. Attachment_1.pdf (73 KB)
   Overview of oral assessments taking place in Switzerland
2. Attachment_2.pdf (80 KB)
   Overview of written examinations taking place in the DACH region
3. Attachment_3.pdf (69 KB)
   Overview of oral examinations taking place in Austria
4. Attachment_4.pdf (70 KB)
   Overview of oral assessments taking place in Germany

References

1. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, et al. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):206-214. DOI: 10.3109/0142159X.2011.551559
2. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Med Educ. 2004;38(2):199-203. DOI: 10.1111/j.1365-2923.2004.01755.x
3. Harden MR, 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
4. van der Vleuten CP. The assessment of professional competence: Developments, research and practical implications. Adv Heal Sci Educ. 1996;1(1):41-67. DOI: 10.1007/BF00596229
5. Fabry G. Medizindidaktik: ein Handbuch für die Praxis. Mannheim: Huber; 2008.
6. 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
7. Downing S. Validity: on the meaning ful interpretation of assessment data. Med Educ. 2003;37(9):830-837. DOI: 10.1046/j.1365-2923.2003.01594.x
8. Ratnapalan S, Hilliard R. Needs Assessment in Postgraduate Medical Education: A Review. Med Educ Online. 2002;7(8):1-8. DOI: 10.3402/meo.v7i.4542

9. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach. 2010;32(8):676-682. DOI: 10.3109/0142159X.2010.500704

10. Bundesärztekammer. (Muster-)Weiterbildungsordnung 2018. Berlin: Bundesärztekammer; 2018. Zugänglich unter/available from: https://www.bundesaerztekammer.de/aerzte/aus-weiter-fortbildung/weiterbildung/muster-weiterbildungsordnung/

11. Frank J, Snell L, Sherbino J. CanMEDS 2015 Physician Competency Framework. Ottawa: Royal College of Physicians and Surgeons of Canada; 2015.

12. Kadmon M, Ganschow P, Gillen S, Hofmann HS, Braune N, Johannink J, Kühn P, Buhr HJ, Berberat PO. Der kompetente Chirurg. Chirurg. 2013;84(10):859-868. DOI: 10.1007/s00104-013-2531-y

13. Jilg S, Möltner A, Berberat P, Fischer MR, Breckwoldt J. How do Supervising Clinicians of a University Hospital and Associated Teaching Hospitals Rate the Relevance of the Key Competencies within the CanMEDS Roles Framework in Respect to Teaching in Clinical Clerkships? GMS Z Med Ausbild. 2015;32(3):Doc33. DOI: 10.3205/zma000975

14. Bürgi H, RindlisbacherB, Bader C, Bloch R, Bosman F, Gasser C, Gerke W, Humair JP, Im Hof V, Kaiser H, Lefebvre D, Schläppi P, Sottas B, Spinas GA, Stuck AE. Swiss Catalogue of Learning Objectives for Undergraduate Medical Training. Genf: Joint Conference of Swiss Medical Faculties (SMIFK); 2007. Zugänglich unter/available from: http://www.smifk.ch

15. Case SM, Swanson DB. Constructing Written Test Questions For the Basic and Clinical Sciences. Philadelphia: National Board of Medical Examiners; 2002. p.112. Zugänglich unter/available from: http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf

16. Schweizerisches Institut für ärztliche Weiter- und Fortbildung. Fortbildung: Investition in die Zukunft. Bern: Schweizer Institut für ärztliche Weiter- und Fortbildung; 2017.

17. Schweizer Institut für ärztliche Weiter- und Fortbildung. Weiterbildungsgänge für weitere sieben Jahre akkreditiert. Bern: Schweizer Institut für ärztliche Weiter- und Fortbildung; 2018.

18. Gerhard-Szep S, Guentsch A, Pospiech P, Soehnel A, Scheutzel P, Wassmann T, Zahn T. Assessment formats in dental medicine: An overview. GMS J Med Educ. 2016;33(4):Doc65. DOI: 10.3205/zma001064

19. Chenot JF, Ehrhardt M. Objective structured clinical examination (OSCE) in der medizinischen Ausbildung: Eine Alternative zur Klausur. Z Allgemeinmed. 2003;79(9):437-442. DOI: 10.1055/s-2003-43064

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

21. Rademakers J, Ten Cate TJ, Bär PR. Progress testing with short answer questions. Med Teach. 2005;27(7):578-582. DOI: 10.1080/01421590500062749

22. Smith S, Kogan JR, Berman NB, Dell MS, Brock DM, Robins LS. The development and preliminary validation of a rubric to assess medical students' written summary statements in virtual patient cases. Acad Med. 2016;91(1):94-100. DOI: 10.1097/ACM.0000000000000800

23. Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: From theory to practice: AMEE Guide No. 75. Med Teach. 2013;35(3):184-193. DOI: 10.3109/0142159X.2013.760036

24. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The Script Concordance Test: A Tool to Assess the Reflective Clinician. Teach Learn Med. 2000;12(4):189-195. DOI: 10.1207/S15328015TLM1204_5

25. Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CP. Script concordance testing: A review of published validity evidence. Med Educ. 2011;45(4):329-338. DOI: 10.1111/j.1365-2923.2010.03863.x

26. Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. Med Educ. 2013;47(12):1175-1183. DOI: 10.1111/medu.12283

27. Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: Insights from a systematic review. Med Educ. 2012;46(6):552-563. DOI: 10.1111/j.1365-2923.2011.04211.x

28. Davis MH, Karunathilake I. The place of the oral examination in today's assessment systems. Med Teach. 2005;27(4):294-297. DOI: 10.1080/01421590500126437

29. Kugler. Mündliche Prüfung Bankfachwirt. In: Schutz A, editor. Mündliche Prüfung Bankfachwirt. Wiesbaden: Gabler; 2007. p.3-6.

30. Memon MA, Joughin GR, Memon B. Oral assessment and postgraduate medical examinations: Establishing conditions for validity, reliability and fairness. Adv Heal Sci Educ. 2010;15(2):277-289. DOI: 10.1007/s10459-008-9111-9

31. Lamping DL. Assessment in health psychology. Can Psychol. 2007;26:121-139. DOI: 10.1037/h0080022

32. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. Acad Med. 1993;68(6):443-451. DOI: 10.1097/00001888-199306000-00002

33. 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

34. Rushforth HE. Objective structured clinical examination (OSCE): Review of literature and implications for nursing education. Nurse Educ Today. 2007;27:481-490. DOI: 10.1016/j.nedt.2006.08.009

35. Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen [Basic quantitative analyses of medical examinations]. GMS Z Med Ausbild. 2006;23(3):Doc53. Zugänglich unter/available from: http://www.egms.de/en/journals/zma/2006-23/zma000272.shtml

36. Hays R. Assessment in medical education: roles for clinical teachers. Clin Teach. 2008;5(1):23-27. DOI: 10.1111/j.1743-498X.2007.00165.x

37. JüngerJ, Just I. Empfehlungender Gesellschaft für Medizinische Ausbildungund des Medizinischen Fakultätentagsfür fakultätsinterne Leistungsnachweise während des Studiums der Human-, Zahn- und Tiermedizin [Recommendations of the GermanSociety for Medical Education and the German Association of Medical Faculties regarding university-specific assessments during the study of human, dental and veterinary medicine]. GMS Z Med Ausbild. 2014;31(3):Doc34. DOI: 10.3205/zma000926

38. Lynch DC, Surdyk PM, Eiser AR. Assessing professionalism: A review of the literature. Med Teach. 2004;26(4):366-373. DOI: 10.1080/01421590410001696434

39. van der Vleuten CP, Verwijnen GM. Fifteen years of experience with progress testing in a problem-based learning curriculum. Med Teach. 1996;18(2):103. DOI: 10.3109/01421599609034142

40. Pell G, Fuller R, Homer M, Roberts T; International Association for Medical Education. How to measure the quality of OSCE: a review of metrics. Med Teach. 2010;32(10):802-811. DOI: 10.3109/0142159X.2010.507716

41. Norcini JJ. Standard setting on educational tests. Med Educ. 2003;37(5):464-449. DOI: 10.1046/j.1365-2923.2003.01495.x

42. Wood TJ, Humphrey-Murto SM, Norman GR. Standard setting in a small scale OSCE: A comparison of the modified borderline-group method and the borderline regression method. Adv Heal Sci Educ Theory Pract. 2006;11(2):115-122. DOI: 10.1007/s10459-005-7853-1

43. Swing SR, Clyman SG, Holmboe ES, Williams RG. Advancing Resident Assessment in Graduate Medical Education. J Grad Med Educ. 2009;1(2):278-286. DOI: 10.4300/JGME-D-09-00010.1

44. Bundesärztekammer. Ärztliche Ausbildung in Deutschland. Weiterbildung. Berlin: Bundesärztekammer; 2015. Zugänglich unter/available from: http://www.bundesaerztekammer.de/aerzte/aus-weiter-fortbildung/ausbildung/allgemeine-informationen-zum-medizinstudium/#c14521

45. Flum E, Maagaard R, Godycki-Cwirko M, Scarborough N, Scherpbier N, Ledig T, Roos M, Steinhäuser J. Assessing family medicine trainees--what can we learn from the European neighbours? GMS Z Med Ausbild. 2015;32(2):Doc21. DOI: 10.3205/zma000963

46. Wass V, van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet. 2001;357(9260):945-949. DOI: 10.1016/S0140-6736(00)04221-5

47. Adler G, von dem Knesebeck J, Hänle MM. Qualität der medizinischen Aus-, Fort- und Weiterbildung. Z Evid Fortbild Qual Gesundhwes. 2008;102(4):235-243. DOI: 10.1016/j.zefq.2008.04.004

48. David DM, Euteneier A, Fischer MR, Hahn EG, Johannink J, Kulike K, Lauch R, Lindhorst E, Noll-Hussong M, Pinilla S, Weih M, Wennekes V. Die Zukunft der ärztlichen Weiterbildung in Deutschland - Positionspapier des Ausschusses Weiterbildung der Gesellschaft für medizinische Ausbildung (GMA). GMS Z Med Ausbild. 2013;30(2):Doc26. DOI: 10.3205/zma000869

49. Driessen E, Scheele F. What is wrong with assessment in postgraduate training? Lessons from clinical practice and educational research. Med Teach. 2013;35(7):569-574. DOI: 10.3109/0142159X.2013.798403

50. Mulder H, Ten Cate O, Daalder R, Berkvens J. Building a competency-based workplace curriculum around entrustable professional activities: The case of physician assistant training. Med Teach. 2010;32(10):e453-459. DOI: 10.3109/0142159X.2010.513719

51. O'Dowd E, Lydon S, O'Connor P, Madden C, Byrne D. A systematic review of 7 years of research on entrustable professional activities in graduate medical education, 2011-2018. Med Educ. 2019;53(3):234-249. DOI: 10.1111/medu.13792

Corresponding author:
Dr. med. Nils Thiessen, MME
EDU - a degree smarter, Digital Education Holdings Ltd., Villa Bighi, Chaplain's House, Kalkara KKR 1320, Republic of Malta
nils.thiessen@edu.edu.mt

Please cite as
Thiessen N, Fischer MR, Huwendiek S. Assessment methods in medical specialist assessments in the DACH region – overview, critical examination and recommendations for further development. GMS J Med Educ. 2019;36(6):Doc78.
DOI: 10.3205/zma001286, URN: urn:nbn:de:0183-zma0012867

This article is freely available from
https://www.egms.de/en/journals/zma/2019-36/zma001286.shtml

Received: *2018-07-29*
Revised: *2019-07-29*
Accepted: *2019-09-04*
Published: *2019-11-15*

Prüfungsmethoden in Facharztprüfungen der DACH-Region – Übersicht, kritische Betrachtung und Empfehlungen zur Weiterentwicklung

Nils Thiessen[1]
Martin R. Fischer[2]
Sören Huwendiek[3]

1 EDU - a degree smarter, Digital Education Holdings Ltd., Kalkara, Republik Malta

2 LMU München, Klinikum der Universität München, Institut für Didaktik und Ausbildungsforschung in der Medizin, München, Deutschland

3 Universität Bern, Institut für Medizinische Lehre, Abteilung für Assessment und Evaluation, Bern, Schweiz

Zusammenfassung

Einleitung: Facharztprüfungen in der Medizin haben die Aufgabe, sicherzustellen, dass Ärzte über die klinischen Kompetenzen verfügen, um ihr Fach eigenverantwortlich zu vertreten und Patienten eigenständig und bestmöglich nach dem aktuellen Stand des Wissens zu versorgen. Bisher gibt es zum Stand der Facharztprüfungen im deutschsprachigen Raum (DACH) keine umfassenden Darstellungen. Deshalb werden in diesem Artikel die zum Einsatz kommenden Prüfungsmethoden in der DACH-Region zusammengestellt, kritisch bewertet und Empfehlungen zur Weiterentwicklung beschrieben.

Methoden: Die Internetseiten der folgenden Institutionen wurden nach Informationen in Bezug auf die eingesetzten Prüfungsmethoden und die Organisation der Facharztprüfungen durchsucht:

1. Homepage des Schweizerischen Instituts für ärztliche Weiter- und Fortbildung (SIWF),
2. Homepage der Akademie der Ärzte (Österreich),
3. Homepage der deutschen Bundesärztekammer (BAEK).

Weiterführende Links wurden berücksichtigt und die erhaltenen Ergebnisse tabellarisch dargestellt. Die in den Facharztprüfungen eingesetzten Prüfungsmethoden werden hinsichtlich etablierter Qualitätskriterien kritisch betrachtet und daraus Empfehlungen zur Weiterentwicklung der Facharztprüfungen abgeleitet.

Ergebnisse: In der Schweiz und in Österreich kommen bereits folgende Prüfungsmethoden zum Einsatz: schriftliche Prüfungen mit Multiple Choice und Kurzantwort-Fragen, strukturierte mündliche Prüfungen, der Script Concordance Test (SCT) und die Objective Structured Clinical Examination (OSCE). Teilweise werden diese Prüfungsmethoden miteinander kombiniert (Triangulation). In Deutschland wird dagegen bisher unstrukturiert mündlich in Form eines "kollegialen Fachgesprächs" geprüft. Damit Wissen, praktische und kommunikative Kompetenzen gleichermaßen überprüft werden können, werden die Methoden-Triangulation und die Beachtung der in diesem Beitrag beschriebenen weiteren Empfehlungen nahegelegt.

Schlussfolgerung: Während es in der Schweiz sowie in Österreich bereits gute Ansätze für qualitätsgesicherte und kompetenzbasierte Facharztprüfungen gibt, ist man in Deutschland noch davon entfernt. Mit den in diesem Artikel vorgestellten Empfehlungen, könnte ein Beitrag dazu geleistet werden, die Facharztprüfungen in der DACH-Region ihren Zielen gemäß zu verbessern.

Schlüsselwörter: Facharztprüfungen, DACH-Region, kognitive, praktische und kommunikative Kompetenzen

Einleitung

Prüfungen haben eine ganze Reihe von wichtigen Aufgaben: Sie besitzen einen stark lernsteigernden Effekt, geben Rückmeldungen über die Wirksamkeit von Aus- und Weiterbildungsprogrammen und schützen den Patienten [1]. Bis Mitte des zwanzigsten Jahrhunderts wurde in der Medizin vornehmlich schriftlich in Form von Aufsätzen oder mündlich geprüft [1], [2]. Aus Prüfungen abgeleitete Bewertungen stellten sich damals häufig als subjektiv, willkürlich und nicht reproduzierbar heraus [1]. Nachfolgend wurden standardisierte Tests wie Multiple Choice Prüfungen (MC-Prüfungen) oder auch Objective Structured Clinical Examinations (OSCE) [3] entwickelt (Case & Swanson 1996, Norcini & Burch 2007, Kogan et al. 2009, zitiert nach Norcini [1]). Prüfungen sollen objektiv, reproduzierbar (reliabel) und valide (gültig) sein. Ferner sollen sie akzeptiert durch Prüflinge und Prüfer sein, eine lernfördernde Komponente haben und möglichst kosteneffizient sein [4]. Mit Objektivität ist gemeint, dass die Prüfung von der Person des Prüfers, von seinen Einstellungen, Gefühlen und Motiven möglichst unabhängig sein soll. Sie bezieht sich auf die Durchführung, Auswertung und die Interpretation einer Prüfung [5]. Eine Prüfung soll bei Wiederholung annähernd das gleiche Ergebnis liefern, also reliabel sein. Die Reliabilität ist ein Maß für die Zuverlässigkeit einer Prüfung. Die Reliabilität wird als Koeffizient von 0 (keine Reliabilität) bis 1 (perfekte Reliabilität) dargestellt. Der Wert 0.80 wird häufig als minimaler Standard für eine bedeutende sog. „high stakes“ Prüfung festgelegt [6]. Eine Prüfung ist u.a. valide, wenn sie das misst, was sie zu messen vorgibt. Sie ist somit ein Maß für die Messgenauigkeit eines Tests [5]. Ein externer Standard zum Vergleich des Prüfungsergebnisses mit diesem Standard wäre für valide Prüfungen wünschenswert, ist aber in der Praxis häufig nicht verfügbar. Dann dienen häufig Experteneinschätzungen für eine Validierung. In der medizinischen Lehre kommen vor allem Konstrukte zur Anwendung, also abstrakte Konzepte und Prinzipien, die aus Verhalten abgeleitet und durch pädagogische und psychologische Theorien erklärt werden [7]. Dieser Sachverhalt wird durch den Begriff der Konstruktvalidität wiedergegeben.

Die Gesellschaft vertraut auf Prüfungen, die gewährleisten, dass Patienten sich in die Obhut von kompetenten und qualifizierten Medizinern geben können, die einen Mindeststandard erreicht haben [1]. Nach Premi sollen Facharztprüfungen sicherstellen, dass Kollegen, die diese Prüfung abgelegt haben, das Wissen und die notwendigen Fertigkeiten ihrer Fachgruppe erworben haben und eigenständig anwenden können (Premi 1994, zitiert nach Ratnapalan & Hilliard [8]). Prüfungen, die das notwendige Wissen sowie die notwendigen Fertigkeiten und Haltungen zur Berufsausübung nachweisen, sind Teil der Selbstregulierung in der medizinischen Weiterbildung. Dies wird weltweit mit zunehmender Skepsis bewertet, erst recht vor dem Hintergrund, dass die Ausbildung zum Arzt eine sehr teure Angelegenheit ist und häufig aus öffentlicher Hand finanziert wird [9]. In Australien, Großbritannien und Canada sind aus diesem Grund die Regierungen direkt mit der Regulierung der medizinischen Weiterbildung betraut (Chantler & Ashton 2009, Shaw et al. 2009, Medicare Advisory Commission 2009, zitiert nach Holmboe [9]).

Die Musterweiterbildungsordnung, die von der BAEK entwickelt wurde, hat empfehlenden Charakter in Bezug auf die Weiterbildung eines Arztes. Der Abschluss der zu dokumentierenden Weiterbildung wird auf Grund der von den Weiterbildungsbefugten erstellten Zeugnisse und einer Prüfung beurteilt. Laut Angaben der BAEK stellt die Weiterbildungsbezeichnung den Nachweis für die erworbene Kompetenz dar und dient als Qualitätssicherung der Patientenversorgung und der Bürgerorientierung. Der Begriff der Kompetenz wird hierbei nicht näher spezifiziert [10].

In der Schweiz ist die FMH (Foederatio Medicorum Helveticorum) der Berufsverband der Schweizer Ärztinnen und Ärzte. Sie ist Dachverband von über 70 Ärzteorganisation. Das Schweizerische Institut für ärztliche Weiter-und Fortbildung (SIWF) ist ein autonomes Organ der FMH und stellt für über 120 Fachgebiete eine qualitativ hochwertige Weiter- und Fortbildung der Ärztinnen und Ärzte sicher. In Zusammenarbeit mit den Fachgesellschaften erlässt das SIWF für jedes Fachgebiet ein detailliertes Weiterbildungs- bzw. Fortbildungsprogramm https://www.siwf.ch/].

Die österreichische Ärztekammer (ÖÄK) verleiht die Berechtigung zur unselbstständigen sowie zur selbstständigen und eigenverantwortlichen ärztlichen Berufsausübung. Für die Durchführung der Arztprüfung als Voraussetzung für die ärztliche Berufsausübung, hat die ÖÄK die Österreichische Akademie für Ärzte GmbH betraut [https://www.aerztekammer.at/]. Die Ausbildungsinhalte und die dazugehörigen Zeugnisse zum Erwerb eines Facharzttitels sind von der ÖÄK erstellt und vorgegeben [https://www.aerztekammer.at/ausbildungsinhalte-und-rasterzeugnisse-kef-und-rz-v-2015].

Die Kompetenzorientierung rückt in der Aus- und Weiterbildung in der Medizin seit einigen Jahren immer mehr in den Vordergrund, dies mit dem Ziel, zu gewährleisten, dass Absolventen die Herausforderungen der praktischen Tätigkeit meistern und alle hierfür notwendigen Kompetenzen besitzen [11].

Das Rollenmodell „Canadian Medical Education Directives for Specialists“ (CanMEDS) wurde von Frank et al. entwickelt, um eine umfassende Weiterbildung von Ärzten zu gewährleisten [12]. Auf der Basis einer systematischen Literaturanalyse sowie breit angelegter Experten- und Stakeholder-Befragungen wurden zur Etablierung des CanMEDS Rahmenmodells durch Frank et al. [11] sieben ärztliche Rollen definiert und in alle Weiterbildungsprogramme Kanadas integriert [12]:

1. Medical Expert,
2. Communicator,
3. Collaborator,
4. Leader,
5. Health Advocate,
6. Scholar,

7. Professional.

Diesen Rollen sind zahlreiche Schlüsselkompetenzen zugeordnet. Das CanMEDS Rollenmodell wurde in Europa bereits in nationalen Lernzielkatalogen für das Medizinstudium integriert (Niederlande (Laan 2010, zitiert nach Jilg) [13], Schweiz [14], Deutschland [http://www.nklm.de]). Nicht nur die medizinische Lehre sollte kompetenzorientiert unterrichtet werden, sondern auch konsekutiv die Weiterbildung. Hierzu gehört unabdingbar auch eine kompetenzorientierte Überprüfung des Wissens, der Fertigkeiten und der Haltungen.
Bisher gibt es keine Zusammenstellung dazu, inwiefern die Facharztprüfungen der DACH-Region kompetenzbasiert gestaltet sind und die anderen genannten Qualitätskriterien für Prüfungen wie Objektivität und Reliabilität Berücksichtigung finden. Ziel dieser Arbeit ist es deshalb, eine Übersicht existierender summativ eingesetzter Facharzt-Prüfungsformate in der DACH-Region und deren Organisation zu erstellen, die Formate kritisch bezüglich Qualitätskriterien zu betrachten und Empfehlungen anhand der internationalen Literatur zu geben. Diese Zusammenstellung soll als ersten Schritt den aktuellen Stand bekannter machen und mögliche Richtungen für die Weiterentwicklung von Facharztprüfungen in der DACH-Region aufzeigen.

Methoden

Folgende Homepages wurden auf Hinweise bestehender Prüfungsformate und die Organisation der Facharztprüfungen durchsucht:

1. Homepage der Schweizerischen Instituts für ärztliche Weiter- und Fortbildung [https://www.siwf.ch/]
2. Homepage der Österreichischen Akademie der Ärzte [https://www.aerztekammer.at/]
3. Homepage der deutschen Bundesärztekammer [10]

Ausgehend von den Webseiten dieser nationalen Dachorganisationen, wurden die Inhalte der Homepages von Fachgesellschaften oder Landesärztekammern, die weiteren Aufschluss über die aktuell eingesetzten Prüfungsformate lieferten, ausgewertet. Die Güte der Internetrecherche ist daher abhängig von den dort gelisteten Informationen und Daten.
Die Durchsicht der einzelnen Geschäftsberichte der Landesärztekammern ergab, soweit verfügbar, einen Überblick über die Anzahl der durchgeführten Prüfungen eines Jahres und den entsprechenden Durchfallquoten. Darüberhinausgehende statistische Kennzahlen oder Kosten wurden nicht zur Verfügung gestellt.
Weiterhin wurden die eingesetzten Prüfungsmethoden zu Qualitätskriterien (Validität, Reliabilität, Objektivität, Akzeptanz, Kosteneffizienz und Einfluss auf das Lernen) von Prüfungen kritisch anhand der international vorliegenden Literatur bewertet und Kriterien für Best Practice Facharztprüfungen aus der Literatur abgeleitet.

Ergebnisse

Übersicht der eingesetzten Prüfungsformate in der DACH-Region

Schweiz

In der Schweiz werden in 26 von 46 Fachgebieten Multiple Choice (MC)-Prüfungen eingesetzt Die Anzahl der Fragen variiert zwischen 50 und 200 Fragen. Die minimale Prüfungsdauer beträgt 120 Minuten, das Maximum liegt bei 360 Minuten. Die Fachgebiete Anästhesiologie, Allergologie/Immunologie sowie Gefäßchirurgie werden in Verbindung mit der European Union of Medical Specialists (UEMS) geprüft. Die Kardiologie plant den Anschluss an die europäische Facharztprüfung für das Jahr 2018. Folgende Fragentypen werden in der Schweiz genutzt, wobei es hier eine deutliche Varianz hinsichtlich des Vorkommens innerhalb der Fachgebiete gibt: Typ A positiv, Typ A negativ, Typ Kprim, Typ B, Typ E, Typ R, Typ Pick N werden als MC-Formate eingesetzt. Zusätzlich werden Short Answer Questions (SAQ) sowie Script Concordance Tests (SCT) genutzt. Eine gute Einführung mit Beispielen zu den einzelnen, unterschiedlichen Fragentypen liefern Swanson und Case [15]. Im Fachgebiet Psychiatrie/Psychotherapie müssen die Prüflinge in einem 2. Teil eine schriftliche Arbeit abgeben und an einem Kolloquium teilnehmen. In der Radiologie u.a. wird mit Freitext Prüfungen geprüft, die nicht näher erläutert werden. Hier steht die Diagnostik von Fällen im Vordergrund. Zum Einsatz kommt ein webbasiertes Prüfungstool. Neben dem schriftlichen Format gibt es in der Schweiz auch mündliche Prüfungen, u.a. Besprechung eines Papers, Vorstellung von Patientendossiers, Durchführung eines Kolloquiums sowie strukturierte mündliche Prüfungen (SMP). Die Dauer variiert von 20 Minuten bis hin zu 180 Minuten. Einige Fachgebiete, z.B. die Endokrinologie/Diabetologie, kombinieren hierbei eine schriftliche mit einer mündlichen Prüfung. In 23 Fachgebieten, also 50% der Facharztprüfungen in der Schweiz, finden Prüfungen mit einem praktischen Anteil statt. Im Fachgebiet der Otho-Rhino-Laryngologie sowie der Thoraxchirurgie wird z.B. praktisch im Rahmen einer Operation geprüft. Die Rheumatologie führt einen 9 Stationen umfassenden OSCE durch, wobei pro Station 10 Minuten zur Verfügung stehen. Hierbei werden nicht nur Wissen (Anatomie, Pathophysiologie u.a.), sondern auch praktische Fertigkeiten (Untersuchungstechniken) sowie kommunikative Fertigkeiten standardisiert überprüft. Im Jahr 2017 wurden durch das SIWF 1428 Facharzttitel [16] und im Jahr 2018 1434 Facharzttitel vergeben [17]. Eine Durchfallquote wird von der SIWF im Geschäftsbericht nicht veröffentlicht.

Österreich

14 von 57 Prüfungen werden mit MC-Fragen durchgeführt. Das Minimum der Fragen wird mit 50 angegeben und reicht bis hin zu 200 Fragen, welche im Fachgebiet

Haut- und Geschlechtskrankheiten zu beantworten sind. Den Prüflingen stehen 60 bis 300 Minuten zur Verfügung. Die Anästhesiologie und Urologie sind der europäischen Facharztprüfung angegliedert. Die Fachgebiete Pathologie und Radiologie setzen Prüfungen mit Kurzantwortfragen ein. Die Dauer umfasst 80-240 Minuten. 45 Fachgebiete prüfen in Österreich mündlich anhand von strukturierten mündlichen Prüfungen (Einsatz eines sog. „Blueprint", vorformulierter Fragen und eines Erwartungshorizontes). Unter „Blueprint" wird in diesem Zusammenhang ein gewichteter inhaltlicher Prüfungsplan verstanden, der sicherstellt, dass eine Auswahl an relevanten Prüfungsinhalten eine inhaltliche Gleichbehandlung eines jeden Prüflings gewährleistet. Bei den meisten Fächern wird ein Blueprint erstellt und explizit bei der Prüfungsbeschreibung erwähnt. Die Dauer der Prüfung kann zwischen 40 und 120 Minuten variieren. Einige Fächer prüfen sowohl schriftlich, als auch mündlich. Eine klinisch-praktische Prüfung gibt es aktuell in Österreich nicht. Eine Statistik über die Anzahl der in Österreich pro Jahr durchgeführten Prüfungen und entsprechender Durchfallquoten liegt uns nicht vor.

Deutschland

Die Facharztprüfung wird bei allen Landesärztekammern in Form einer unstrukturierten mündlichen Prüfung (UMP) abgehalten, die mindestens 30 Minuten dauert und bis zu 60 Minuten dauern kann. Diese Prüfungsform wird bei allen Facharztbezeichnungen eingesetzt und auch als „kollegiales Fachgespräch" bezeichnet. Die Anzahl der Prüfer kann variieren. Mindestens ein Prüfer muss aus dem zu prüfenden Fachgebiet stammen. Die Prüfungsergebnisse sind zu dokumentieren. Typischerweise wird weder ein strukturierter Blueprint erstellt, noch werden im Vorfeld im Sinne einer standardisierten und strukturierten Prüfung Fragen vorformuliert und ein Erwartungshorizont angegeben. Die Landesärztekammer Hamburg stellt hier eine Ausnahme dar. Hier werden die Fragen im Vorfeld an den Prüfungsvorsitzenden überreicht. Tabelle 1 zeigt die Anzahl der Facharztprüfungen und dazugehörige Durchfallquoten. Die Daten wurden denjenigen Geschäftsberichten des Jahres 2017 der jeweiligen Landesärztekammern, die online verfügbar waren, entnommen. Eine Anfrage an die BAEK hinsichtlich einer umfassenden, bundesweiten Statistik lieferte das Ergebnis, dass eine solche Statistik nicht verfügbar sei.

Kritische Würdigung der in der DACH-Region eingesetzten Prüfungen

MC-Prüfungen

MC-Prüfungen sind im Bereich der Medizin als Prüfungsmethode weit verbreitet, da sie kosteneffizient einsetzbar sind und für das Prüfen von Wissen eine hohe Validität und Reliabilität bieten können (Norcini 1985, zitiert nach Gerhard-Szep [18]). Dies setzt aber voraus, dass inhaltlich und formal hochwertige Fragen in ausreichender Anzahl pro Prüfung (mindestens 40) eingesetzt werden (Jünger 2014, zitiert nach Gerhard-Szep [18]). Case et al. betonen, dass zwei Kriterien zur Entwicklung einer guten Frage notwendig sind: Die Frage muss relevanten Inhalt prüfen und gut strukturiert sein [15]. Die MC-Fragen-Entwicklung auf qualitativ hohem Niveau ist zeitaufwendig. Mit schriftlichen Prüfungsmethoden ist es vor allem möglich, Faktenwissen zu prüfen. Kommunikative und praktische Fertigkeiten bzw. Kompetenzen mittels MC-Fragen zu prüfen ist, im Gegensatz zu einer OSCE, nicht möglich [19].

Kurzantwortfragen

Bei Kurzantwortfragen müssen frei formulierte, kurze, stichwortartige Antworten gegeben werden. Prüfungsteilnehmer müssen spontan über die richtige Lösung nachdenken und können nicht auf vorgegebene Antwortmöglichkeiten reagieren [5]. Dies vermindert das sog. „Cueing", welches Prüflingen die Möglichkeit gibt, eine Frage ohne Wissen richtig zu beantworten (Schuwirth 2004, zitiert nach Epstein [20]). Idealerweise werden auch hier „kontextreiche" Fragenstämme (Fallvignetten) angeboten, die auch das Prüfen von Anwendungswissen und z.B. „Clinical Reasoning" ermöglichen. Die Reliabilität hängt wesentlich auch von der Qualität der Bewertungen durch die Prüfer ab [20] – hier kann ein Training der Prüfer im Vorfeld Abhilfe schaffen. Die Auswertung ist anfälliger für subjektive Verzerrungen als bei MC-Fragen. Vorformulierte Erwartungshorizonte, an denen sich die auswertenden Prüfer orientieren müssen, können die Objektivität erhöhen und sollten vorliegen. Akzeptable Reliabilitätswerte erreicht man durch den Einsatz mehrerer Prüfer, die jeweils für die Auswertung unterschiedlicher Aufgaben zuständig sind [5]. Eine anschauliche Vorlage einer Aufgabenstellung liefern Rademakers et al. [21]. Zwischenzeitlich gibt es auch die Möglichkeit, die Antwortoptionen computerbasiert auszuwerten [22]. Hier sind in naher Zukunft neue Entwicklungen zu erwarten, die auf Methoden der künstlichen Intelligenz zurückgreifen.

Script Concordance Test (SCT)

Der SCT dient zur Überprüfung der „Clinical Reasoning" Kompetenz von Prüflingen in Situationen klinischer Unsicherheit [23]. Es werden kurze klinische Szenarien geschildert und schrittweise weitere Zusatzinformationen gegeben. Der Prüfling soll nun im Lichte dieser neuen Informationen diagnostische, weiterführende oder therapeutische Entscheidungen treffen [24]. Anhand einer 5 Punkte umfassenden Likert Skala von -2 bis +2 muss der Prüfling angeben, inwiefern die Zusatzinformation die Krankheitshypothese, die im Szenario beschrieben ist, unterstützt oder eben nicht [25]. Die Ergebnisse der Prüflinge werden im Nachgang mit den Einschätzungen einer Expertengruppe abgeglichen; dabei erzielt die Antwort die meisten Punkte, bei der die meisten Experten zugestimmt haben (Goldstandard) [23]. Abbildung 1 zeigt ein Beispiel mit drei Fragen [25].

Tabelle 1: Übersicht der Prüfungskennzahlen mach Geschäftsberichten der Landesärztekammern für das Jahr 2017

	Durchgeführte Prüfungen	Bestandene Prüfungen	Durchfallquote in %
Ärztekammer Berlin	1256	1182	6
Ärztekammer Nordrhein	1731	1651	4,6
Ärztekammer Hamburg	513	501	2,3
Ärztekammer Brandenburg	322	299	7,1
Ärztekammer Bremen	127	122	3,9
Ärztekammer des Saarlandes	185	179	3,2
Ärztekammer Sachsen-Anhalt	314	298	5
Ärztekammer Westfalen-Lippe	1288	1182	8,2
			5,0375

Sie denken an einen/eine	Anschließend erhalten Sie folgende Zusatzinformation:	Ihre Hypothese ist nun*				
1. Zerebralen Abszess	Der Patient hatte eine Mittelohrentzündung vor 10 Tagen	-2	-1	0	+1	+2
2. Ischämischen Schlaganfall	Plötzliches Einsetzen der Symptomatik vor 2 h	-2	-1	0	+1	+2
3. Zerebrale Metastase	Normales CCT mit Kontrastmittel	-2	-1	0	+1	+2

*-2=hinfällig, -1=unwahrscheinlich, 0=weder wahrscheinlicher, noch unwahrscheinlicher, +1=wahrscheinlicher, +2=sicher zutreffen
CCT=Craniales Compoutertomogramm

Abbildung 1: Fall 1: Sie beurteilen einen 75jährigen Mann mit einer Hemiparese rechts. Sie befinden sich im Notfallzentrum

Zwischenzeitlich konnten verschiedene Arbeitsgruppen günstige psychometrische Eigenschaften des SCT nachweisen (Konstruktvalidität, Reliabilität und Durchführbarkeit) [23]. Brailovsky et al. (2001, zitiert nach Epstein [20]) konnten zeigen, dass die Antworten auf solche Fragen mit dem Ausbildungsstand des Prüflings korrelieren und dessen künftige Leistung bei mündlichen Prüfungen in Bezug auf dessen Fähigkeit des „clinical reasoning“ vorhersagen [20]. Kritisch ist, dass eine 5 Punkte Likert Skala zu Missverständnissen und Angabe falscher Bewertungen seitens des Expertenpanels führen kann, so dass Lineberry et al. die Verwendung einer 3 Punkte Skala bestehend aus „widerlegt“, „weder widerlegt, noch unterstützt“ und „zutreffend“ empfehlen. Zusätzlich besteht die Gefahr, dass Prüflinge ihre Antworten im Sinne der Tendenz zur Mitte angeben und damit ein besseres Prüfungsergebnis erhalten, als diejenigen, die die Likert Skala in ihren Extremen ausnutzen [26]. Zudem ist weiterhin in Diskussion, wie sinnvoll die Punktevergabe - einer Expertengruppe entsprechend – ist, zumal hierfür 10-20 Mitglieder [27] empfohlen werden und der SCT damit recht aufwändig ist.

Strukturierte Mündliche Prüfung

Die mündliche Prüfung ist eine traditionelle Form der Prüfung, bei der ein oder mehrere Prüfer Fragen an den Kandidaten richten. Die mündliche Prüfung soll das Wissen bewerten, die Tiefe des Wissens erforschen und andere Qualitäten wie die geistige Beweglichkeit testen. Colton & Peterson, Foster et al. und Kelly et al. kritisierten den Einsatz von mündlichen Prüfungen in High-Stakes-Prüfungen aufgrund ihrer geringen Reliabilität (Peterson 1967, Foster et al. 1969, Kelly et al. 1971, zitiert nach Davis [28]. Bei der mündlichen Prüfung treten zahlreiche Fehlerquellen auf, denen Prüfer im Rahmen unterliegen. Beispielsweise dominiert beim Primacy-Effekt der erste Eindruck über spätere, beim Recency-Effekt setzen sich spätere Eindrücke nachhaltiger fest. Beim Halo-Effekt überstrahlt die Wahrnehmung und Bewertung einer Eigenschaft, die Wahrnehmung und Bewertung anderer Eigenschaften. Antipathie, Sympathie und die Zusammensetzung der Prüfer haben ebenso einen Einfluss auf die Bewertung der Prüfungsleistung [29]. Nach Roloff et al. steigen Reliabilität und Objektivität, wenn mehrere Prüfer unabhängig voneinander prüfen und die Anzahl der Fragen sowie die Prüfungszeit zunehmen (Roloff 2016, zitiert nach Gerhard-Szep [18]). Memon et al. haben 15 Qualitätssicherungsmaßnahmen benannt, die aus der Sicht der Literatur notwendig sind, um Objektivität, Reliabilität und Validität von Facharztprüfungen zu gewährleisten. Sie geben an, dass die mündliche Prüfung am ehesten geeignet ist „Clinical Reasoning“ und „Decision Making“ zu prüfen. Der Prüfungsinhalt sollte durch ein Expertengremium zuvor festgelegt werden. Die Prüfungsfragen sollten dahingehend ausgewählt werden, dass sie nicht nur die entsprechende Wissenstiefe, sondern auch die Breite des Stoffgebietes adäquat prüfen und eine entsprechende inter-item Reliabilität gewährleisten. Prüfer müssen zuvor in Bezug auf eine mündliche Prüfung geschult werden. Abweichungen zwischen Prüfern (inter-examiner variations) müssen überwacht und adressiert werden. Item-Erstellungs- und Implementierungsprozesse müssen standardisiert sein und eine statistische Auswertung soll Rückschlüsse auf die Reliabilität geben. Bei mündlichen Prüfungen muss in der Bewertung mit Verzerrungen (Bias)

gerechnet werden und deshalb sollte dahingehend eine Qualitätssicherung durchgeführt werden [30].

Unstrukturierte Mündliche Prüfung

Die unstrukturiert mündliche Prüfung sieht zumeist die Anzahl von zwei ungeschulten Prüfern vor, die basierend auf ihren Prüfungserfahrungen, prüfen. Hierbei gibt es typischerweise weder einen vorformulierten Erwartungshorizont, noch zuvor verschriftliche Fragen, die anhand des Curriculums bzw. Blueprint ausgerichtet sind. Bereits 1985 konnten Jayawickramarajah et al. darstellen, dass 2/3 der Fragen im Rahmen einer unstrukturierten mündlichen Prüfung ausschließlich Faktenwissen prüften. Zusätzliches Problem einer unstrukturierten mündlichen Prüfung ist die hohe Wahrscheinlichkeit des Auftretens der Konstrukt Irrelevanten Varianz (KIV), dadurch, dass eine zu geringe Anzahl an Prüfern eingesetzt wird. KIV tritt auf, wenn beispielsweise die zu prüfende Kompetenz „Clinical decision making“ durch Auftreten, Angst, Sprachfertigkeit oder Kleidung des Prüflings beeinflusst werden. Konstrukt Unterrepräsentation (KU) ist eine weitere Hürde, die im Rahmen einer unstrukturiert mündlichen Prüfung beachtet werden muss, da zum Beispiel zwei bis drei klinische Szenarien, die geprüft werden, nicht die ganze Breite des zu prüfenden Stoffgebietes abdecken können. Bedenken hinsichtlich der Validität dieser traditionellen Prüfungsform haben dazu geführt, dass diese Prüfungsform durch schriftliche Prüfungen oder durch strukturierte mündliche Prüfungen ersetzt wurden (Jayawickramarajah et al. 1985, Turnball, Danoff, & Norman, 1996, Pokorny & Frazier, 1966, zitiert nach Lamping et al. 2007 [31]).

OSCE

Das OSCE-Prüfungsformat wurde in den 70iger Jahren von Harden entwickelt und prüft vor allem klinisch-praktische Kompetenzen. Durch Standardisierung wird eine höhere Objektivität erreicht [3]. Damit eine standardisierte Darstellung von Erkrankungen erreicht werden kann, werden Schauspieler speziell für diesen Einsatz trainiert [32]. Eine Reihe von Problemstellungen wird dem Prüfling in Form eines Parcours präsentiert. Die Anzahl sowohl der Stationen, als auch der Prüfer, wirkt sich positiv auf die Reliabilität aus. Trotz hoher Varianz in Studien, kann bei mehr als 10 Stationen bereits eine gute Reliabilität erreicht werden [33]. Pro Station hat der Prüfling ca. 5-15 Minuten Zeit, um die Aufgabenstellung zu bearbeiten [2]. Der Prüfer überprüft die gesehene klinische Kompetenz anhand einer Checkliste und/oder einer globalen Ratingskala. OSCEs machen es möglich, dass im Rahmen des Kontakts mit dem standardisierten Patienten (SP) durch geschickte Anamnesetechniken und eine patientenzentrierte körperliche Untersuchung, eine Diagnose gestellt werden kann. Trainierte SPs können nach Van der Vleuten und Tamblyn nicht von richtigen Patienten unterschieden werden, wiederholt zuverlässig spielen und auch wertvolles Feedback an die Prüflinge geben (Van der Vleuten 1990, Tamblyn 1991, zitiert nach Newble [2]). Die OSCE wird von Studierenden und Dozierenden im Allgemeinen positiv betrachtet (Roberts & Brown 1990, zitiert nach Rushfort [34]) auch, wenn die Studierenden die Prüfung z.T. als Stress empfinden. Der Prüfung wird im Vergleich zu anderen Prüfungen eine größere Objektivität bescheinigt (Schuwirth & Van der Vleuten 2003, zitiert nach Rushforth [34]) und hat einen positiven Effekt auf die Motivation (Bartfey et al. 2004, zitiert nach Rushforth [34]) der Prüflinge hierfür zu lernen [34].

Überblick der Bewertung der Prüfungsformate

Die nachfolgende Tabelle 2, in Anlehnung an Gerhard-Szep et al., Lubarsky et al., Epstein et al. und Van der Vleuten et al. [4], [18], [20], [25] wurde konzipiert, um die in der DACH-Region auftretenden Prüfungsformen zu vergegenwärtigen und aus Sicht der Autoren hinsichtlich beschriebener, wesentlicher Qualitätskriterien einzuordnen. Sie dient im Wesentlichen der besseren Übersicht und soll die Empfehlungen zu einer Best Practice Facharztprüfung unterstützen.

Empfehlung einer Best Practice Facharztprüfung

Im Folgenden werden Empfehlungen für eine Best Practice Facharztprüfung gegeben, die der aktuellen Literatur abgeleitet wurden.
Die Beachtung der folgenden Empfehlungen unterstützt, dass die resultierenden Prüfungen möglichst valide, reliabel, objektiv, akzeptiert, lehrreich und kosteneffizient sind. Dies ist notwendig, damit kompetenzorientierte Lernziele sinnvoll geprüft werden können und Prüfungen belegen, dass Prüflinge die zur eigenständigen Behandlung von Patienten notwendigen Kompetenzen erlernt haben.

1. **Einsatz verschiedener Prüfungsmethoden (Triangulation):** Unterschiedliche Prüfungsmethoden (z.B. eine schriftliche (e-)Prüfung für Wissen und eine OSCE-Prüfung für praktische und kommunikative Fertigkeiten) sollten genutzt werden, um einerseits Wissen und andererseits praktische und kommunikative Fertigkeiten adäquat prüfen zu können. Erst die Kombination (Triangulation) der Ergebnisse verschiedener Prüfungsformate kann eine hohe Validität und unterschiedliche Kompetenzen sicherstellen [35].
2. **Vorab Festlegung der zu prüfenden Inhalte und Kompetenzen (Blueprinting):** Ein gewichteter Prüfungsplan (sog. Blueprint) gibt den Rahmen für die Prüfung vor, indem er vor einer Prüfung sicherstellt, dass eine ausgewogene Auswahl relevanter Lernziele in die Prüfung einfließt [36]. So soll gewährleistet werden, dass die Prüfung inhaltlich gültig, fair, relevant und für das geprüfte Fach repräsentativ ist.
3. **Vorab Festlegung der Fragestellungen und des Erwartungshorizontes:** Bei mündlichen und praktischen Prüfungen müssen genauso wie bei schriftlichen

Tabelle 2: Darstellung der in der DACH–Region verwandten Prüfungsmethoden hinsichtlich relevanter Merkmale, von „++“ (=hoch) bis „–“ (=niedrig) bzw. geeignet bis ungeeignet aus der Sicht der Autoren.

	Objektivität	Reliabilität*	Aufwand	Prüfung von Wissen	Prüfung von praktischen Fertigkeiten
MC - Prüfungen	++	++	+	++	--
SMP	++	++	++	++	--
OSCE	++	++	++	+	++
SCT	++	++	+	++	--
SAQ	++	++	+	++	--
UMP	--	--	--	-	--

* = abhängig von der Anzahl und Qualität der Prüfungsaufgaben

Formaten die Fragestellungen und der Erwartungshorizont im Vorfeld pro Frage/Station jeweils schriftlich festgehalten werden (sog. Strukturierung). Klar strukturierte Checklisten stellen bei mündlichen und praktischen Prüfungen den Erwartungshorizont unmissverständlich dar und sorgen damit für die notwendige Interpretations- und auch Auswertungs-Objektivität [18], [37].

4. **Ausreichende Anzahl an Fragen, Prüfern, Stationen/zu prüfender Lernziele:** Für relevante Prüfungen wird eine Mindestreliabilität von 0,8 angegeben [4], [35]. Um diese zu verbessern, kann die Anzahl der Aufgaben und/oder deren Qualität gesteigert werden [35]. Ebenso hat die Anzahl der Prüfer einen steigernden Effekt auf die Reliabilität (Swanson 1987, zitiert nach Lynch [38]). Je mehr Prüfer prüfen, desto besser wird die Reliabilität. In mündlichen und praktischen Prüfungen ist es sinnvoller, je einen Prüfer pro Thema/Station zu haben, anstatt mehrere Prüfer gleichzeitig und dafür weniger Stationen/Themen.
5. **Qualitätssicherung der erstellten Fragen/Aufgaben:** Der inhaltliche und formal-sprachliche Review und die Revision der Aufgabenstellungen sind zur Gewährleistung der Eindeutigkeit der Lösungen und hohen Qualität der Aufgaben und Fragen notwendig. Die Validität der Prüfungsergebnisse wird durch einen Reviewprozess (medizindidaktisch geschulte Experten) der Fragen und Aufgaben gestärkt [4], [39].
6. **Qualitätssicherung bei der Auswertung der Prüfung:** Die Qualitätssicherung durch teststatistische Auswertung von Prüfungen ermöglicht, OSCE-Stationen und Prüfungsaufgaben gezielt zu überarbeiten, Checklisten zu prüfen und ggf. auch Schlussfolgerung auf die Qualität des Unterrichts zu ziehen. Als Parameter werden folgende empfohlen: Bei schriftlichen Prüfungen sollte zumindest eine Untersuchung bzgl. Reliabilität, Trennschärfe und Item-Schwierigkeit durchgeführt werden (außer bei kleinen Kandidatenzahlen <30, wegen des Einflusses des Zufalls). Bei mündlichen bzw. praktischen Prüfungen ist die Untersuchung der Trennschärfe und Item-Schwierigkeit und zusätzlich der Reliabilität auf Posten- und Item-Niveau, besser „OSCE-Metrics“ [40] sehr wünschenswert. Ein modernerer Ansatz zur Festlegung der Bestehensgrenze von Prüfungen, bei der ein Bestehen auch bei mehr oder weniger als 50% gelöster Aufgaben erfolgen kann, verfolgt den Ansatz der inhaltlichen Festlegung der Bestehensgrenze [36], (z.B. modifiziertes Angoff-Verfahren bei MC [41], Borderline Regression Methode bei OSCE (siehe Wood et al. [42]).
7. **Lerneffekt für die Prüflinge:** Prüfungen dienen nicht nur zur Entscheidungsfindung, sondern sind ebenso ein sehr wichtiger Lernanreiz für die Kandidaten und unterstützen zudem durch die Feedbackgabe von Prüfungsergebnissen an Prüflinge zusätzlich den Lerneffekt [4]. So kann z.B. ein Feedbackbrief so gestaltet sein, dass die Prüflinge wissen, in welchen Bereichen des Blueprints sie weniger gut im Vergleich zu den anderen Aufgaben und zu den anderen Prüflingen abgeschnitten haben.

Berücksichtigung der Kosteneffizienz: Gute Prüfungen haben ihren Preis, stellen aber definitiv eine lohnenswerte Investition im Hinblick auf den Lerneffekt von Prüflingen dar [4], [39]. Es sollte jeweils die Prüfungsmethode gewählt werden, die einerseits den Prüfungsgegenstand adäquat prüfen kann (Inhaltsvalidität) und andererseits dabei noch möglichst kosteneffizient ist. Wenn es also z.B. primär um das Überprüfen von Anwendungswissen bei vielen Kandidaten geht, ist eine schriftliche Prüfung mit Vignettenfragen einer strukturierten mündlichen Prüfung in Sachen Kosteneffizienz überlegen. Der Zusammenschluss von Fachgesellschaften, kann den Aufwand z.B. für praktische Prüfungen (z.B. OSCE) mit dem Ziel der Überprüfung der CanMEDS-Rollen reduzieren (vgl. dem Schweizer Basisexamen in der Chirurgie im Bereich Wissen https://basisexamen.ch/).

Diskussion

Diese Arbeit geht der Frage nach, welche Facharztmethoden in der DACH-Region eingesetzt werden. Zudem werden die eingesetzten Prüfungsmethoden kritisch betrachtet und basierend auf der aktuellen Literatur Empfehlungen zur Weiterentwicklung der Facharztprüfungen beschrieben.

Eingesetzte Prüfungsmethoden

Mehr als 50% der Facharztprüfungen werden in der Schweiz in Form von MC-Prüfungen durchgeführt. Die Fachgebiete Anästhesiologie, Allergologie/Immunologie sowie Gefäßchirurgie werden in Verbindung mit der European Union of Medical Specialists (UEMS) geprüft. Weitere Fachbereiche planen dies. Sieben verschiedene MC-Fragen Typen werden eingesetzt, ebenso der SAQ, Freitextprüfungen (nicht näher spezifiziert) sowie der SQT. Auch schriftliche Arbeiten, die SMP, praktische Prüfungen und eine OSCE werden genutzt. Insgesamt haben 50% der Schweizer Facharztprüfungen einen praktischen Anteil (vgl. Anhang 1 und Anhang 2).

In Österreich werden 25% der Facharztprüfungen als schriftliche Prüfungen (MC-Fragen) durchgeführt. Zwei Fachgebiete prüfen in Verbindung mit der UEMS. SAQ und SMP sind ebenfalls verwandte Prüfungsformen. Ein Blueprinting wird regelhaft eingesetzt. Eine praktische Prüfung gibt es bislang nicht (vgl. Anhang 2 und Anhang 3).

In Deutschland findet flächendeckend eine UMP statt, die als „kollegiales Fachgespräch“ bezeichnet wird (vgl. Anhang 4).

Kritische Betrachtung und Empfehlungen zu Weiterentwicklung der Facharztprüfungen

Positiv zu bewerten ist, dass in der Schweiz bereits 50% der Facharztprüfungen einen praktischen Prüfungsanteil haben. Um im Rahmen der Facharztprüfung auch praktische und kommunikative Kompetenzen prüfen zu können, sollte eine praktische, Kommunikations- und Kompetenzorientierte Prüfung, zusätzlich zu einer wissensorientierten Prüfungsmethode (schriftlich/SMP), eingesetzt werden. Das OSCE-Format könnte hier zum Einsatz kommen. Zur Kostenreduktion könnten z.B. zumindest Anteile der Prüfungen landesweit durchgeführt werden.

Zudem ist anhand der Literatur positiv zu werten, dass in der Schweiz und in Österreich MC-Prüfungen mit Typ A pos. Fragen in der Mehrzahl genutzt werden, um Wissen inkl. Anwendungswissen objektiv zu überprüfen und teststatistisch auszuwerten. Eine Beurteilung der überprüften Kompetenz anhand dieser Prüfungsformen ist jedoch nur dann möglich, wenn ein detaillierter Einblick in die durchgeführten Prüfungen und deren Ergebnisse ermöglicht werden kann. Die Vorbereitung von schriftlichen Prüfungen wird häufig unterschätzt und ist zeitaufwendig, da sowohl inhaltliche als auch formal-sprachliche und (medizin-)didaktische Review-Prozesse stattfinden müssen, um die Eindeutigkeit der Lösungen zu garantieren. Der Arbeitsaufwand wird hierbei vor allem in die Erstellungsphase verlagert (vgl. Gerhard-Szep [18]). Zusätzlich werden ethische und kulturelle Fragestellungen bei der Fragenerstellung häufig vermieden, da kontextreiche Fragen schwierig zu schreiben sind (Frederiksen 1984, zitiert nach Epstein) [20]. Swing et al. empfehlen hier generell den Einsatz von regelmäßigen Prüferschulungen sowie den Einsatz von Expertengruppen, die die eingesetzte Prüfungsmethode regelmäßig kritisch hinterfragen [43]. Anwendungswissen kann nicht nur schriftlich, sondern auch strukturiert mündlich geprüft werden. Neben Anwendungswissen kann eine SMP klinische Entscheidungsfindung, professionelles Denken, Selbstvertrauen und Selbstbewusstsein bewerten [30]. Zu Bedenken ist, dass jede SMP durch den hohen Raum- und Personalbedarf mit hohen Kosten verbunden ist. Blueprinting und die Erstellung eines Erwartungshorizontes sind ebenfalls erforderlich. Dauer, Anzahl und Erfahrung der Prüfer haben einen direkten Einfluss auf Gütekriterien der SMP (Roloff 2016, zitiert nach Gerhard-Szep) [18]. Eine Prüferschulung kann psychologische Fehlerquellen (siehe Kugler 2007 [29]), denen Prüfer unbewusst unterliegen, bewusstmachen und reduzieren helfen. Hier muss mit Widerstand der bisherigen Prüfer gerechnet werden, die jahrelang geprüft haben, ohne dahingehend geschult worden zu sein.

In Deutschland wird unstrukturiert mündlich geprüft, obwohl die Bundesärztekammer ausdrücklich auf ihrer Webseite hervorhebt, dass die Weiterbildungsbezeichnung als Nachweis für die erworbene Kompetenz und der Qualitätssicherung der Patientenversorgung und der Bürgerorientierung dient [44]. Diesem Anspruch werden die aktuellen Facharztprüfungen in Deutschland also nicht gerecht, da UMP nicht die erforderlichen Qualitätskriterien erfüllen (vgl. auch Tabelle 2). Man kann daher nur vermuten, dass die UMP, einer Tradition folgend, entwickelt und bislang keiner kritischen Überprüfung unterzogen wurde. Sie UMP kann jedoch nicht empfohlen werden, wenn Landesärztekammern ihrer Verantwortung für die medizinische Qualitätssicherung zukünftig besser gerecht werden wollen. Ein erster Schritt könnte darin liegen, dass sie mit medizinischen Fakultäten in Kontakt treten, welche bereits seit einigen Jahren Erfahrungen in Bezug auf die Implementierung von strukturierten Prüfungsmethoden gesammelt haben. Auch könnte ein Erfahrungsaustausch mit Experten aus der DACH-Region erfolgen, die bereits eine Qualitätssicherung der eingesetzten Facharztprüfungen durchführen. Um unstrukturiert mündliche Prüfungen kurzfristig in strukturiert mündliche Prüfungen umzuwandeln, könnten in Deutschland die Landesärztekammern medizindidaktisch geschulte Kollegen bitten, Prüferschulungen vor Ort durchzuführen. Flum betont den Aspekt der Qualitätssicherung ebenfalls, indem sie sagt, dass es, zur Sicherstellung der Behandlungsqualität und Patientensicherheit, hilfreich wäre, Kompetenzen und Prüfungsmethoden in der Weiterbildung Allgemeinmedizin innerhalb der EU zu standardisieren [45]. Ebenso sollen Facharztprüfungen

als ein Instrument zur Regulation der Inhalte der Weiterbildung angesehen werden und dem tatsächlichen Versorgungsbedarf der Bevölkerung sowie den Lernzielen und dem Curriculum gerecht werden [45].
Eine Empfehlung in Bezug auf eine Best Practice Facharztprüfung erfolgt vor dem Hintergrund, dass bislang nur vereinzelt Artikel zu Prüfungsmethoden im Bereich der Weiterbildung Stellung bezogen. Ein kombinierter Einsatz von Prüfungsmethoden ist unerlässlich (Triangulation), um das notwendige Kompetenzspektrum abdecken zu können. Ebenfalls sollte ein obligates Blueprinting (Dauphinee 1994, zitiert nach Wass), die Erstellung von Fragen und des Erwartungshorizontes vorab, eine ausreichende Anzahl an Fragen/Aufgaben, Maßnahmen der Qualitätssicherung in Bezug auf Erstellung und Auswertung von Prüfungsfragen/Aufgaben, der Einsatz von Prüfungsfeedback (Gronlund 1998, zitiert nach Norcini) sowie der Einsatz möglichst kosteneffizienter Prüfungsmethoden eingesetzt und dokumentiert werden. Unterstützt werden diese Empfehlungen durch zahlreiche Publikationen [1], [37], [46]. Caraccio et al. betonen, dass verschiedene Prüfungsformen zu kombinieren seien, damit das Kompetenzniveau von Weiterbildungsassistenten eingeschätzt werden kann (Caraccio 2013, zitiert nach Flum [45]). Taylor et al. weisen auf eine notwendige Standardisierung von Facharztprüfungen hin, deren Kosten allerdings berücksichtigt werden müssten (Taylor 1998, zitiert nach Flum [45]). David et al. und Adler et al. diskutieren an dieser Stelle zu Recht, dass Kosten, die im Rahmen der Weiterbildung entstehen, durch das DRG System abgebildet sein müssten [47], [48]. Der notwendige Einsatz von Blueprinting im Bereich der postgraduierten Weiterbildung wird durch Wass et al. unterstützt. Um den Lernerfolg bereits während der Weiterbildung zu optimieren, sollten Kompetenzen auch kontinuierlich anhand verschiedener Methoden erfasst und Feedback anhand formativer Prüfungen gegeben werden (z.B. Mini Clinical Examination (Mini-CEX), Direct Observation of Procedural Skills (DOPS), Portfolios etc.) [1], [49]. Kompetenzbasierte Curricula sollten berücksichtigen, wie Wissen, Fertigkeiten und Haltungen auf der höchsten Prüfungsebene „does“ nach Miller (Miller 1990) geprüft werden [50]. Prüfungen sollten kompetenzbasiert sein [11]. Aktuell werden zunehmend sogenannte EPAs (Entrustable Professional Activities) genutzt, um die Curriculumsentwicklung zu unterstützen und neue Wege des kompentenzbasierten Lernens und Prüfens zu erproben [51]. Eine solche professionelle Aktivität (EPA) könnte beispielsweise das Erkennen eines Notfallpatienten auf der Normalstation und die initiale Beurteilung und das Einleiten notwendiger medizinischer Maßnahmen sein. Eine Besonderheit des EPA-Ansatzes ist die Beurteilung des Lernenden anhand der vermuteten Supervisionsnotwendigkeit („Entrustment“). Der Einsatz von EPAs ist für klinisch tätige Ärzte häufig intuitiv attraktiv, allerdings muss das Potential inkl. der bestehenden Herausforderungen (vgl. auch die Literatur zu arbeitsplatzbasierten Assessments weiter untersucht werden [51], bevor ein Ersatz von qualitätsgesicherten summativen Facharztprüfungen dadurch erwogen werden könnte.

Limitiert ist diese Arbeit zum einen sicherlich dadurch, dass die im Internet recherchierten Daten nur deskriptiv dargestellt werden können. Zum anderen sind Weiterbildungscurricula, die Durchführung der Weiterbildung sowie die Gestaltung von Prüfungen eng miteinander im Sinne des „Constructive Alignment“ verzahnt. Eine weitere Limitation ist daher, dass wir in diesem Artikel vornehmlich auf die Darstellung der eingesetzten Prüfungen in der DACH – Region eingegangen sind und Weiterbildungscurricula nur am Rande erwähnt werden konnten. Ein Folgeartikel könnte jedoch die unterschiedlichen Weiterbildungscurricula und deren Umsetzung innerhalb der DACH-Region erläutern.

Schlussfolgerung

In der DACH-Region sind die Organisation von Facharztprüfungen, die eingesetzten Prüfungsmethoden und Qualitätssicherungsmaßnahmen sehr unterschiedlich. In Österreich und der Schweiz kommen - im Gegensatz zu Deutschland - bereits strukturierte und standardisierte Facharztprüfungsmethoden zum Einsatz, in der Schweiz auch praktische Prüfungen. Wenn Facharztprüfungen sicherstellen sollen, dass die Fachärzte über die notwendigen Kompetenzen zur Patientenversorgung verfügen, müssten diese auch derart gestaltet sein, dass sie Kompetenzen auch tatsächlich abprüfen können. Dies erscheint aktuell in allen Fachbereichen in Deutschland, aber auch in den meisten Fachbereichen in Österreich und der Schweiz, noch nicht der Fall zu sein. Daher ist es notwendig, um die Qualität der Weiterbildung zu sichern, dass in den drei Ländern noch mehr darauf geachtet wird, dass eine summative Facharztprüfung die vorgesehenen Kompetenzen der prospektiven Fachärzte auch überprüft. Empfohlen wird derzeit eine Kombination einer schriftlichen Prüfung mit einer praktischen Prüfung (z.B. OSCE), da auf diese Weise nicht nur Wissen, sondern auch weitere Kompetenzen inkl. praktischer und kommunikativer Fertigkeiten geprüft werden können (vgl. Tabelle 2).

Danksagung

Wir danken Frau Dr. med. Susanne Frankenhauser, MME, Frau Dr. med. Uta Krämer und Herrn Brian Webber für die konstruktive Hilfe bei der Durchsicht des Manuskriptes.

Anmerkung

Geschlechtergerechte Sprache: Aus Gründen der besseren Lesbarkeit wird nachfolgend die männliche Anrede verwendet. Sämtliche Personenbezeichnungen gelten gleichermaßen für beide Geschlechter.

Interessenkonflikt

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

Anhänge

Verfügbar unter
https://www.egms.de/de/journals/zma/2019-36/zma001286.shtml

1. Anhang_1.pdf (73 KB)
   Übersicht über stattfindende Mündliche – Prüfungen in der Schweiz
2. Anhang_2.pdf (79 KB)
   Übersicht über stattfindende schriftliche Prüfungen in der DACH-Region
3. Anhang_3.pdf (72 KB)
   Übersicht über stattfindende Mündliche - Prüfungen in Österreich
4. Anhang_4.pdf (70 KB)
   Übersicht über stattfindende Mündliche – Prüfungen in Deutschland

Literatur

1. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, et al. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):206-214. DOI: 10.3109/0142159X.2011.551559
2. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Med Educ. 2004;38(2):199-203. DOI: 10.1111/j.1365-2923.2004.01755.x
3. Harden MR, 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
4. van der Vleuten CP. The assessment of professional competence: Developments, research and practical implications. Adv Heal Sci Educ. 1996;1(1):41-67. DOI: 10.1007/BF00596229
5. Fabry G. Medizindidaktik: ein Handbuch für die Praxis. Mannheim: Huber; 2008.
6. 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
7. Downing S. Validity: on the meaning ful interpretation of assessment data. Med Educ. 2003;37(9):830-837. DOI: 10.1046/j.1365-2923.2003.01594.x
8. Ratnapalan S, Hilliard R. Needs Assessment in Postgraduate Medical Education: A Review. Med Educ Online. 2002;7(8):1-8. DOI: 10.3402/meo.v7i.4542
9. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach. 2010;32(8):676-682. DOI: 10.3109/0142159X.2010.500704
10. Bundesärztekammer. (Muster-)Weiterbildungsordnung 2018. Berlin: Bundesärztekammer; 2018. Zugänglich unter/available from: https://www.bundesaerztekammer.de/aerzte/aus-weiter-fortbildung/weiterbildung/muster-weiterbildungsordnung/
11. Frank J, Snell L, Sherbino J. CanMEDS 2015 Physician Competency Framework. Ottawa: Royal College of Physicians and Surgeons of Canada; 2015.
12. Kadmon M, Ganschow P, Gillen S, Hofmann HS, Braune N, Johannink J, Kühn P, Buhr HJ, Berberat PO. Der kompetente Chirurg. Chirurg. 2013;84(10):859-868. DOI: 10.1007/s00104-013-2531-y
13. Jilg S, Möltner A, Berberat P, Fischer MR, Breckwoldt J. How do Supervising Clinicians of a University Hospital and Associated Teaching Hospitals Rate the Relevance of the Key Competencies within the CanMEDS Roles Framework in Respect to Teaching in Clinical Clerkships? GMS Z Med Ausbild. 2015;32(3):Doc33. DOI: 10.3205/zma000975
14. Bürgi H, RindlisbacherB, Bader C, Bloch R, Bosman F, Gasser C, Gerke W, Humair JP, Im Hof V, Kaiser H, Lefebvre D, Schläppi P, Sottas B, Spinas GA, Stuck AE. Swiss Catalogue of Learning Objectives for Undergraduate Medical Training. Genf: Joint Conference of Swiss Medical Faculties (SMIFK); 2007. Zugänglich unter/available from: http://www.smifk.ch
15. Case SM, Swanson DB. Constructing Written Test Questions For the Basic and Clinical Sciences. Philadelphia: National Board of Medical Examiners; 2002. p.112. Zugänglich unter/available from: http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf
16. Schweizerisches Institut für ärztliche Weiter- und Fortbildung. Fortbildung: Investition in die Zukunft. Bern: Schweizer Institut für ärztliche Weiter- und Fortbildung; 2017.
17. Schweizer Institut für ärztliche Weiter- und Fortbildung. Weiterbildungsgänge für weitere sieben Jahre akkreditiert. Bern: Schweizer Institut für ärztliche Weiter- und Fortbildung; 2018.
18. Gerhard-Szep S, Guentsch A, Pospiech P, Soehnel A, Scheutzel P, Wassmann T, Zahn T. Assessment formats in dental medicine: An overview. GMS J Med Educ. 2016;33(4):Doc65. DOI: 10.3205/zma001064
19. Chenot JF, Ehrhardt M. Objective structured clinical examination (OSCE) in der medizinischen Ausbildung: Eine Alternative zur Klausur. Z Allgemeinmed. 2003;79(9):437-442. DOI: 10.1055/s-2003-43064
20. Epstein RM. Medical education - Assessment in medical education. N Engl J Med. 2007;356(4):387-396. DOI: 10.1056/NEJMra054784
21. Rademakers J, Ten Cate TJ, Bär PR. Progress testing with short answer questions. Med Teach. 2005;27(7):578-582. DOI: 10.1080/01421590500062749
22. Smith S, Kogan JR, Berman NB, Dell MS, Brock DM, Robins LS. The development and preliminary validation of a rubric to assess medical students' written summary statements in virtual patient cases. Acad Med. 2016;91(1):94-100. DOI: 10.1097/ACM.0000000000000800
23. Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: From theory to practice: AMEE Guide No. 75. Med Teach. 2013;35(3):184-193. DOI: 10.3109/0142159X.2013.760036
24. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The Script Concordance Test: A Tool to Assess the Reflective Clinician. Teach Learn Med. 2000;12(4):189-195. DOI: 10.1207/S15328015TLM1204_5
25. Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CP. Script concordance testing: A review of published validity evidence. Med Educ. 2011;45(4):329-338. DOI: 10.1111/j.1365-2923.2010.03863.x
26. Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. Med Educ. 2013;47(12):1175-1183. DOI: 10.1111/medu.12283

27. Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: Insights from a systematic review. Med Educ. 2012;46(6):552-563. DOI: 10.1111/j.1365-2923.2011.04211.x

28. Davis MH, Karunathilake I. The place of the oral examination in today's assessment systems. Med Teach. 2005;27(4):294-297. DOI: 10.1080/01421590500126437

29. Kugler. Mündliche Prüfung Bankfachwirt. In: Schutz A, editor. Mündliche Prüfung Bankfachwirt. Wiesbaden: Gabler; 2007. p.3-6.

30. Memon MA, Joughin GR, Memon B. Oral assessment and postgraduate medical examinations: Establishing conditions for validity, reliability and fairness. Adv Heal Sci Educ. 2010;15(2):277-289. DOI: 10.1007/s10459-008-9111-9

31. Lamping DL. Assessment in health psychology. Can Psychol. 2007;26:121-139. DOI: 10.1037/h0080022

32. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. Acad Med. 1993;68(6):443-451. DOI: 10.1097/00001888-199306000-00002

33. 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

34. Rushforth HE. Objective structured clinical examination (OSCE): Review of literature and implications for nursing education. Nurse Educ Today. 2007;27:481-490. DOI: 10.1016/j.nedt.2006.08.009

35. Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen [Basic quantitative analyses of medical examinations]. GMS Z Med Ausbild. 2006;23(3):Doc53. Zugänglich unter/available from: http://www.egms.de/en/journals/zma/2006-23/zma000272.shtml

36. Hays R. Assessment in medical education: roles for clinical teachers. Clin Teach. 2008;5(1):23-27. DOI: 10.1111/j.1743-498X.2007.00165.x

37. JüngerJ, Just I. Empfehlungender Gesellschaft für Medizinische Ausbildungund des Medizinischen Fakultätentagsfür fakultätsinterne Leistungsnachweise während des Studiums der Human-, Zahn- und Tiermedizin [Recommendations of the GermanSociety for Medical Education and the German Association of Medical Faculties regarding university-specific assessments during the study of human, dental and veterinary medicine]. GMS Z Med Ausbild. 2014;31(3):Doc34. DOI: 10.3205/zma000926

38. Lynch DC, Surdyk PM, Eiser AR. Assessing professionalism: A review of the literature. Med Teach. 2004;26(4):366-373. DOI: 10.1080/01421590410001696434

39. van der Vleuten CP, Verwijnen GM. Fifteen years of experience with progress testing in a problem-based learning curriculum. Med Teach. 1996;18(2):103. DOI: 10.3109/01421599609034142

40. Pell G, Fuller R, Homer M, Roberts T; International Association for Medical Education. How to measure the quality of OSCE: a review of metrics. Med Teach. 2010;32(10):802-811. DOI: 10.3109/0142159X.2010.507716

41. Norcini JJ. Standard setting on educational tests. Med Educ. 2003;37(5):464-449. DOI: 10.1046/j.1365-2923.2003.01495.x

42. Wood TJ, Humphrey-Murto SM, Norman GR. Standard setting in a small scale OSCE: A comparison of the modified borderline-group method and the borderline regression method. Adv Heal Sci Educ Theory Pract. 2006;11(2):115-122. DOI: 10.1007/s10459-005-7853-1

43. Swing SR, Clyman SG, Holmboe ES, Williams RG. Advancing Resident Assessment in Graduate Medical Education. J Grad Med Educ. 2009;1(2):278-286. DOI: 10.4300/JGME-D-09-00010.1

44. Bundesärztekammer. Ärztliche Ausbildung in Deutschland. Weiterbildung. Berlin: Bundesärztekammer; 2015. Zugänglich unter/available from: http://www.bundesaerztekammer.de/aerzte/aus-weiter-fortbildung/ausbildung/allgemeine-informationen-zum-medizinstudium/#c14521

45. Flum E, Maagaard R, Godycki-Cwirko M, Scarborough N, Scherpbier N, Ledig T, Roos M, Steinhäuser J. Assessing family medicine trainees–what can we learn from the European neighbours? GMS Z Med Ausbild. 2015;32(2):Doc21. DOI: 10.3205/zma000963

46. Wass V, van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet. 2001;357(9260):945-949. DOI: 10.1016/S0140-6736(00)04221-5

47. Adler G, von dem Knesebeck J, Hänle MM. Qualität der medizinischen Aus-, Fort- und Weiterbildung. Z Evid Fortbild Qual Gesundhwes. 2008;102(4):235-243. DOI: 10.1016/j.zefq.2008.04.004

48. David DM, Euteneier A, Fischer MR, Hahn EG, Johannink J, Kulike K, Lauch R, Lindhorst E, Noll-Hussong M, Pinilla S, Weih M, Wennekes V. Die Zukunft der ärztlichen Weiterbildung in Deutschland - Positionspapier des Ausschusses Weiterbildung der Gesellschaft für medizinische Ausbildung (GMA). GMS Z Med Ausbild. 2013;30(2):Doc26. DOI: 10.3205/zma000869

49. Driessen E, Scheele F. What is wrong with assessment in postgraduate training? Lessons from clinical practice and educational research. Med Teach. 2013;35(7):569-574. DOI: 10.3109/0142159X.2013.798403

50. Mulder H, Ten Cate O, Daalder R, Berkvens J. Building a competency-based workplace curriculum around entrustable professional activities: The case of physician assistant training. Med Teach. 2010;32(10):e453-459. DOI: 10.3109/0142159X.2010.513719

51. O'Dowd E, Lydon S, O'Connor P, Madden C, Byrne D. A systematic review of 7 years of research on entrustable professional activities in graduate medical education, 2011-2018. Med Educ. 2019;53(3):234-249. DOI: 10.1111/medu.13792

Korrespondenzadresse:

Dr. med. Nils Thiessen, MME
EDU - a degree smarter, Digital Education Holdings Ltd., Villa Bighi, Chaplain's House, Kalkara KKR 1320, Republik Malta
nils.thiessen@edu.edu.mt

Bitte zitieren als
Thiessen N, Fischer MR, Huwendiek S. Assessment methods in medical specialist assessments in the DACH region – overview, critical examination and recommendations for further development. GMS J Med Educ. 2019;36(6):Doc78.
DOI: 10.3205/zma001286, URN: urn:nbn:de:0183-zma0012867

Artikel online frei zugänglich unter
https://www.egms.de/en/journals/zma/2019-36/zma001286.shtml

Eingereicht: *29.07.2018*
Überarbeitet: *29.07.2019*
Angenommen: *04.09.2019*
Veröffentlicht: *15.11.2019*