

Essential steps in the development, implementation, evaluation and quality assurance of the written part of the Swiss federal licensing examination for human medicine

Abstract

Purpose: This report describes the essential steps in the development, implementation, evaluation and quality assurance of the written part of the Swiss Federal Licensing Examination for Human Medicine (FLE) and the insights gained since its introduction in 2011.

Methods: Based on existing scientific evidence, international expertise, and experience gained from previous examinations, the FLE is developed by experts from all five medical faculties in Switzerland with the support of the Institute for Medical Education and is held simultaneously at five locations. The exam organisers document and review every examination held and continuously optimise the processes; they have summarised the results in this report.

Results: The essential steps comprise the development, revision and translation of questions; construction of the exam and production of materials; candidate preparation; implementation and analysis. The quality assurance measures consist of guideline coherence in the development of the questions and implementation of the exam, revision processes, construction of the exam based on the national blueprint, multiphase review of the translations and exam material, and statistical analysis of the exam and the comments from candidates.

The intensive collaboration, especially on the part of representatives from all the participating faculties and a central coordination unit, which provides methodological support throughout and oversees the analysis of the exam, has proven successful. Successfully completed examinations and reliable results in the eleven examinations so far implemented represent the outcomes of the quality assurance measures. Significant insights in recent years are the importance of appreciating the work of those involved and the central organisation of exam development, thus ensuring the long-term success of the process.

Conclusion: Common guidelines and workshops, quality assurance measures accompanied by the continuous improvement of all processes, and appreciation of everyone involved, are essential to carrying out such an examination at a high-quality level in the long term.

Keywords: national final examination, licensing examination, summative assessment, multiple choice

Tina Schurter¹
Monica Escher²
David Gachoud³
Piotr Bednarski^{4,5}
Balthasar Hug^{6,7}
Roger Kropf^{6,8}
Juliane
Meng-Hentschel¹
Benjamin König¹
Christine Beyeler¹
Sissel Guttormsen¹
Sören Huwendiek¹

1 University of Bern, Institute for Medical Education, Department for Assessment and Evaluation, Bern, Switzerland

2 University of Geneva, Medical Faculty, Geneva, Switzerland

3 University of Lausanne, Medical Faculty, Lausanne, Switzerland

4 University of Fribourg, Medical Faculty, Fribourg, Switzerland

5 University of Bern, Medical Faculty, Bern, Switzerland

6 University of Basel, Medical Faculty, Basel, Switzerland

7 University of Lucerne, Medical Faculty, Lucerne, Switzerland

8 University of Zurich, Medical Faculty, Zurich, Switzerland

1. Introduction

As a result of the Federal Act of 23 June 2006 on University Medical Professions (Medical Professions Act, LPMed), which came into effect on 1 September 2007, the Swiss Federal Licensing Examination for Human Medicine (FLE) had to be restructured and centrally organised. Following a preparation period, accordingly, a national (federal) licensing examination was introduced in 2011 which is to be taken after completion of medical studies at master's level based on current learning objectives (2011-2020: *Swiss Catalogue of Learning Objectives for Undergraduate Medical Training (SCLO)* [1], since 2021 *Principal Relevant Objectives and Framework for Integrative Learning and Education in Switzerland (PROFILES)* [<http://www.profilesmed.ch/>]). On successful completion, the candidates may practise their medical profession under supervision and commence their further medical training. The examination consists of two parts [2]: a multiple-choice exam (MC) and a structured practical clinical exam involving standardised patients (clinical skills (CS)) [3]. This has replaced the former final examinations which were done in several specialties. The challenge in developing this new examination was to design – in the most efficient way possible and by involving the five existing medical faculties in two language regions – a common cross-disciplinary, application-orientated exam weighted according to a multidimensional blueprint that would guarantee eligibility for further training at the national level, meet international quality standards, and be legally robust.

There is considerable interest in national licensing examinations [4], [5], [6]. Other countries are considering their introduction (Great Britain [6] and Norway) or looking to harmonise the practical exam in situations where a common written licensing examination already exists (Germany [7]). There are few publications, however, on experience and insights from the joint development of cross-faculty or national written examinations [8]. Edwards et al. [9] demonstrated, in the Australian context, that the following factors in particular foster successful collaboration on exam development: *committed group* (e.g. *medical schools, stakeholders*), *funding* (e.g. *development grant*), *engagement* (*team meetings to build vision*), *products* (e.g. *framework, assessment development*), *ownership* (e.g. *open to all medical schools*). In addition, articles have been published on specific aspects such as automated question generation for the national exam in Canada [8], [10] or the use of computer-based [11], [12] or adaptive [13] tests in such a framework.

Overall, there are only few publications on how to develop a high-quality written exam designed specifically as a national licensing examination. Considering the additional aspect of bilingual implementation, even fewer reports can be found.

This report follows up on previous publications concerning the Swiss FLE [2], [3]. It addresses the essential steps, quality assurance measures and insights from the perspective of the organisers of the written part of the FLE

following the eleven exam cycles that have taken place to date and offers interested parties the chance to compare and obtain reference points for similar projects of their own.

2. Methods

Setting

The MC exam is an examination with two parts, each of which consists of 150 MC questions to be completed in 4.5 hours. It is closely coordinated and takes place simultaneously at all five medical faculties. For general information on this examination, please refer to the previously published overview article [2]. In addition to subject matter experts from the hospitals, representatives from general medical practice, methodology experts (expertise in question development, revision and evaluation), IT specialists, administrators, professional translators, representatives from the Federal Office of Public Health (FOPH), and the federal board of examiners, are involved in the examination process. The board of examiners is the most important decision-making body and is made up of vice-deans for teaching from the medical faculties and representatives from general medical practice, continuing education, and the FOPH. The methodology experts make up the national MC working group with representatives from the faculties and the Institute for Medical Education (IML). Except for recruiting question writers and implementing the examination at the faculties, the IML is responsible for coordination.

Development of this project report

This project report is based on the experience gained from the eleven written federal examinations held since 2011. After each examination season, the recent exam along with quality assurance and quality indices are reviewed and documented to ensure continuation of what has proven effective and to identify potential areas for improvement. The essential development, implementation, evaluation and quality assurance steps, along with the insights, are collated by the exam organisers in an iterative process.

3. Results

The examination process is explained in detail below. Figure 1 illustrates the most important steps of the process; table 1 provides a summary of quality assurance measures and insights.

3.1. Question development and revision

Around 60 experts nominated by the faculties write draft questions that are discussed and, if necessary, revised in author groups (peer review) during the first day of a two-day workshop. On the second day of the workshop,

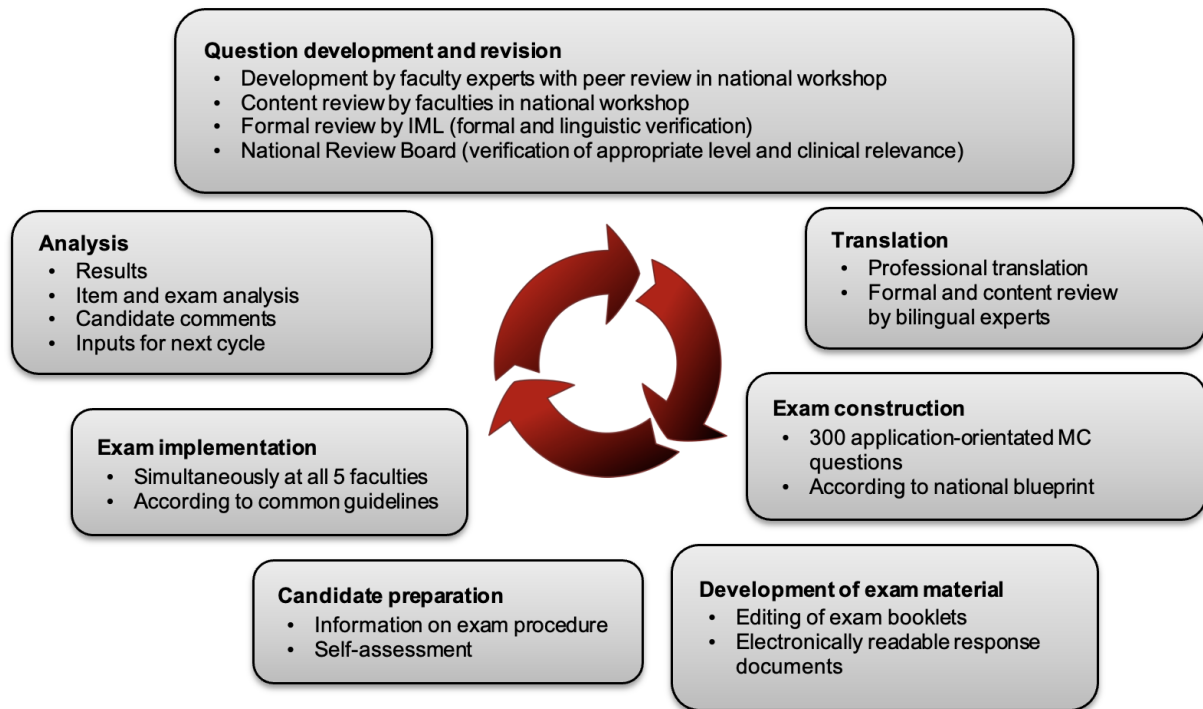


Figure 1: Essential steps in the examination process

the new questions are reviewed for a second time and finalised in editorial groups consisting of at least one representative from each faculty and representatives from different medical disciplines. Any old questions that were created five years previously or were conspicuous in the most recent examination due to statistical deviations or candidate comments are reviewed for their relevance and revised by the same editorial groups. Following a centralised formal and linguistic review of the new questions by methodology experts at the IML, the questions are examined by a national review board (consisting of physicians from different disciplines of all faculties who are well versed in the guidelines on question development for this examination and also write questions themselves) to verify their suitability (level of difficulty) and clinical relevance before they are approved in the database for construction of the next exam. The questions are managed using software developed at the IML [<https://www.iml.unibe.ch/themen/uebersichten/projekte>].

Quality assurance

The requirements for the question writers cover the content-related basics (learning objectives and the resulting content blueprint), information on the types of questions, the specified standard structure for case vignettes, and formulation principles according to Krebs [14]. Preference is given to questions with case vignettes and thus application-orientated questions involving, among other things, diagnostic or therapeutic decisions.

The guidelines published by the Medical Professions Commission (MEBEKO) state the following regarding the relevance of the questions to practical application [15]:

“The clinical knowledge exam (written) tests the interdisciplinary, application-orientated knowledge of the entire spectrum of problems in human medicine across the disciplines” and “The questions should relate to a specific problem, wherever possible, that is presented in a case or problem vignette”. This is specified in detail, with examples, in the internal instructions or training documents for the authors. Each question should be as authentic as possible, i.e. reflect a doctor’s daily work. To this end, it should refer wherever possible to a specific case in which a patient is described (age, gender, setting, reason for consultation, medical history, status/findings, possible results from diagnostic tests, perhaps initial treatment, follow-up results etc.) and a specific question must be answered, e.g. “What is the most likely diagnosis?”, “What is the most useful therapeutic approach?”. In addition, every question should be clinically relevant.

With respect to the question types, MEBEKO provides the following definition [15]: “Selection of the only correct or best from 3-5 offered choice answers (type A, positively or negatively formulated) and fourfold decision right/wrong (type Kprim).”

The questions to be written are determined based on the learning objectives and blueprint (see table 2). The blueprint also defines the percentage distribution of the questions in terms of content. All authors contribute also to teaching at their faculties and developing questions for the faculty exams. As part of the workshop, they receive informational material beforehand concerning the development of draft questions that will familiarise them with the FLE standards. The involvement of all faculties in every revision step (including deletion of questions in exceptional cases) facilitates the development of a na-

Table 1: Summary of quality assurance and insights

Steps	Quality assurance	Insights
Question development and revision	<ul style="list-style-type: none"> • Common guidelines on question development • Group work with individual preparation • Inclusion of all faculties in all revision steps for both content (considering among other things authenticity and clinical relevance of the questions) & formal review • Survey of workshop participants 	A two-day joint workshop leads to greater output than numerous half-day workshops in small (local) groups and is perceived by the experts as respectful and motivating.
Translation	<ul style="list-style-type: none"> • Multiphase translation review • Exclusion of incorrectly translated questions from the assessment 	An additional review of the translation by a medical professional in their native language improves the translations.
Exam construction	<ul style="list-style-type: none"> • Compliance with the blueprint • Selection of questions previously used following good performance • Utilisation of findings from surveying candidates after the exam 	To achieve a searchable database, it is essential to tag the questions. The analysis data can also be searched and filtered.
Development of exam material	<ul style="list-style-type: none"> • Multiphase review of examination booklets 	A review is necessary, as mistakes are occasionally detected.
Candidate preparation	<ul style="list-style-type: none"> • Annual verification and adaptation of the information • Development and maintenance of an online self-assessment tool to prepare formally for the respective level and formats • Informational events for the candidates at the faculties • Analysis of objections to the exam result 	Not all candidates actively study the information on the examination if the material is only available online.
Exam implementation	<ul style="list-style-type: none"> • Common guidelines on exam implementation • Utilisation of findings from surveying candidates after the exam 	Some candidates find it difficult to absorb information or instructions on the day of the exam.
Analysis	<ul style="list-style-type: none"> • Survey of candidates after the exam and, among others, analysis of comments on deficiencies in individual exam questions • Four-eyes principle for the most important steps • Exclusion of exam questions with formal or content-related deficiencies from the scoring and feedback to the authors and review board concerning the deficiencies • Item statistic <i>discrimination</i> • Item statistic <i>difficulty</i> • Reliability • Definition of passing score 	Reporting of previous deficiencies has helped to continuously improve the question development process; hence, the number of questions having to be excluded from the evaluation due to formal or content-related deficiencies has continually decreased over the years.

tionally valid examination accepted by all training institutes.

Insights

The survey of the workshop participants revealed that the format used since 2017 for developing questions (workshop without daily distractions, exchange with expert colleagues throughout Switzerland) is regarded as highly motivating [16], unlike the decentralised process of writing questions alone that was in place until 2016. The workshop format also increases the output of new questions of higher quality (less need for revision). This con-

centrated two-day effort is facilitated by the secure web-based question database mentioned above.

3.2. Translation

Depending on the author, the questions are formulated in German or French. Following the revision process, they are translated into the other language, respectively, by professional translators. The translations undergo formal and linguistic review by methodology experts, while medical specialists check the content.

Table 2: Current blueprint (BP) (according to PROFILES, valid as of 2021) for the written part of the Swiss Federal Licensing Examination for Human Medicine – main dimensions 1 and 2

Dimension 1: Situations as Starting Points (SSP)	Dimension 2: Medical Task		
	1. Understanding mechanisms of disease	2. Assessment/ Diagnosis	3. Management
1. General complaints and symptoms (SSP 1-16)			
2.1 Head and neck (SSP 17-39)			
2.2 Chest (SSP 40-51)			
2.3 Abdomen (SSP 52-61)			
2.4 Pelvis, urogenital system (SSP 62-80)			
2.5 Back and extremities (SSP 81-86)			
2.6 Skin (SSP 87-95)			
2.7 Nervous system (SSP 96-105)			
2.8 Injuries and trauma (SSP 106-117)			
2.9 Emotional and behavioural symptoms (SSP 118-130)			
3.1 Findings upon physical examination (SSP 131-149)			
3.2 Findings upon additional examination (SSP 150-174)			
4.1 Situations related to pregnancy and motherhood (SSP 175-183)			
4.2 Situations related to childhood (SSP 184-195)			
4.3 Situations related to old age (SSP 196-202)			
4.4 Emergency symptoms, findings and situations (SSP 203-220)			
4.5 Issues linked with prevention and health promotion (SSP 221-227)			
4.6 Palliative care (SSP 228-232)			
4.7 Psychosocial issues (SSP 233-241)			
4.8 Various health care issues (SSP 242-265)			

Quality assurance

The quality of the translations is assured by multiphase review. After the examination, candidate comments and the patterns of responses to the questions are analysed to identify any translation errors or technical ambiguities. Questions containing translation errors are excluded from the evaluation, as are other questions with possible formal deficiencies.

Insights

Isolated translation issues have been addressed in some candidate comments. Hence, in addition to the existing control process, all questions now undergo a review by a medical professional in their native language. Having adapted and introduced additional control steps in the translation review process, fewer questions have had to be rejected due to poor translations.

3.3. Exam construction

A total of 300 questions are selected from the question database for each examination in line with the nationally applicable blueprint (for the first two dimensions, see table 2), which was developed on the basis of the learning

objectives [1], [<http://www.profilesmed.ch/>]. The current blueprint consists of three main dimensions:

- Dimension 1: *Situations as starting points*,
- Dimension 2: *Medical tasks*, and
- Dimension 3: *General objectives* (focus on medical expert). *Type of condition* (*acute, subacute, chronic*) and *setting* (*ambulatory practice, hospital, nursing home for elderly people, other*) are further dimensions of the blueprint.

To prevent the strategic omission of individual fields of knowledge during study, the percentage target values are not published and thus are not specified in the table. At least 20% of the questions that have proven effective in previous examinations are used. The compatibility of learning objectives and exam questions is guaranteed by the clear definitions provided in the examination regulations, and content-wise by the blueprint and the criteria for the types of questions selected.

Quality assurance

The representative distribution of the content is ensured by the blueprint and by the candidate survey following the exam using a questionnaire with open-ended and closed-ended questions. By tagging the questions in the

question database, compliance with the blueprint is supported.

Insights

Analysis of the questionnaire revealed that it is important to many candidates to cover the studied subjects as comprehensively as possible (comments on open-ended questions in the questionnaire) and to have questions that in terms of content reflect the learning objectives (with a median of 4 on a scale of 1 (do not agree at all) to 5 (completely agree), closed-ended question in the questionnaire). Compliance with the blueprint and broad coverage is only possible thanks to detailed tagging of the questions.

3.4. Development of exam material

The templates for the exam booklets are generated from the question database and printed before being sent to the faculties. The subjects covered by the questions are divided equally between the two exam booklets; the length of the text in each booklet is identical.

Quality assurance

The quality of the exam booklets is assured by means of multiphase review involving verification of the content and formal aspects (form and language) by experts.

Insights

Despite the preceding revision process during question development and translation, anomalies are still occasionally detected when reviewing the print proofs. These are checked and corrected by physicians during a content review phase.

3.5. Candidate preparation

Candidates can obtain information from the website of the Federal Office of Public Health and faculty informational events concerning the examination procedure [15] and receive nationally standardised documents from the person responsible at the respective site. In addition, around 300 representative sample questions are made available online in a self-assessment tool [<https://www.iml.unibe.ch/angebot/assessment/pruefungsdienstleistungen/self-assessment>].

Quality assurance

The preparatory information is reviewed every year and adapted, as necessary.

Insights

Feedback from the candidates in the questionnaire and argumentation in objections to the exam result have shown that not all candidates actively study the informa-

tion provided prior to the exam. Hence, this information is also sent directly to each candidate in advance.

3.6. Exam implementation

The exam takes place on two days at the same time (4.5 hours each) at five faculties. The exam is taken not only by successful graduates of master's programmes in Switzerland but also by candidates holding unrecognised foreign medical qualifications issued outside of the EU/EFTA.

Quality assurance

Common guidelines on exam implementation regulate the conditions of the examination premises, aids permitted, and instructions to be read out during the exam, for example.

Insights


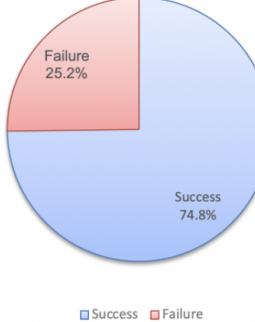
Some candidates find it difficult to absorb information or instructions on the day of the exam. Hence, they additionally receive all the information in advance. Suggestions entered in the questionnaire by the candidates help to continually improve the procedure.

3.7. Evaluation and results

An item analysis is performed for every exam question set (assessment of item measurement properties). Difficulty and discrimination are primarily assessed; these remain visible in the question pool. If a question is reused, it can thus be compared across different exam years. The aim is to achieve a minimum of 0.2 as the item discrimination index (r), and 50-90% for P as the degree of difficulty. The multiple true/false questions of the Kprim type are analysed using the half-point method developed at the IML [17].

The exam results and candidate comments on questions that appear to be flawed are initially analysed by methodology experts. Conspicuous questions are discussed with several clinical experts and excluded from the analysis if formal or content-related deficiencies are detected: for example, if a question that is too difficult is found to be of "specialist" level, or a previously overlooked formal error is identified when the r value is too low.

The passing score is based on two internationally recognised standard-setting methods (Angoff and Hofstee [18], [19], [20]) and an analysis based on the Rasch model [21]. The Rasch model is the simplest model of the item response theory (IRT) in which the item difficulties and the ability of the candidates can be estimated as a means of explaining the exam results. By using items with a difficulty already observed in previous examinations that are therefore known (referred to as anchor items), the model permits the pass requirement to be maintained at a consistent level over the years even if the difficulty of the exam varies each year. Put simply: a comparison between anchor items and the rest of the examination

Results 2020	Faculty candidates**	Candidates with an unrecognised foreign qualification***
Total number of candidates	1105	103
Mean score (percentage of points achieved)	76.6%	60.6%
Standard deviation	6.9%	11.5%
Success/failure rate		

* The 2020 results are comparable with all results since 2011.

** Faculty candidates: candidates having completed their studies at a Swiss faculty and sitting the exam for the first time.

*** Candidates with an unrecognised foreign qualification: candidates having completed their studies outside Switzerland and the EU/EFTA and sitting the exam for the first time.

Figure 2: Exam results 2020*

means the examination difficulty can be judged against the previous years, and the passing score can thus be adapted accordingly. The pass grade is therefore set by the board of examiners using the standard-setting methods and the Rasch model.

Quality assurance

The final decision concerning the passing score is made by the national board of examiners. Formal or content-related deficiencies in exam questions discovered during the analysis are reported to the authors and the review board so that the next round of question development can be optimised accordingly. Questions with poor difficulty and discrimination scores and content-related deficiencies are not reused. All important analysis steps are carried out under the four-eyes principle: prior to publication, the most important results are reviewed by a second individual based on a checklist. For example: Was the passing score ultimately agreed by the board of examiners and recorded in the meeting minutes actually applied in the results lists, the charts in the analysis report, and in the letters to the candidates? Are the answers on the answer sheets of candidates who have failed actually consistent with the electronically recorded answers considered in the results calculation? The analyses represent an essential final step in assuring the quality of the entire examination process; further details are provided in Chapter 4 of [14].

Insights

Continuously optimised quality assurance measures throughout the examination process, including reporting

of deficiencies discovered in the exam questions to the authors and the review board, mean that the number of questions having to be excluded from the analysis due to formal or content-related deficiencies has continually decreased over the years (approx. 40 from the first (13.3%) and about 15 (5%) from the most recent examination). The results from the eleven examinations implemented to date revealed thoroughly reliable exam results (Cronbach's alpha average 0.90, range 0.87-0.91). Figure 2 lists the results from the examination 2020 as an example. The success rates of the candidates who had completed their studies at a Swiss faculty (referred to as "faculty candidates") were consistently high, at 99.5%, whereas those of candidates with an unrecognised foreign medical qualification were much lower. This could be interpreted in such a way that Swiss candidates have also been taught using the Swiss Catalogue of Learning Objectives and thus there is a good alignment between their curriculum and examination, whereas this is not the case with candidates from abroad. Another explanation could be the linguistic challenges faced by foreign candidates or that their qualification has often been obtained several years previously and their level of knowledge is no longer the same.

4. Discussion

Important quality assurance measures entail, among others, the common guidelines on the development of questions and implementation of the examination, the preparation of questions in national workshops, the multiphase revision process involving all faculties, construction of the examination based on the national blue-

print, and multiphase review of the translations and the exam material. Analysis of the examination, including analysis of candidate comments, is a final, significant quality assurance step in the entire examination process. Though Edwards et al. [9] also described some of these processes, including success factors, we can add insights into aspects of question development in national workshops, bilingualism, and analysis. Furthermore, the inclusion of all faculties in the design of the blueprint and in all phases of question development, revision and implementation leads to greater acceptance among the faculties and increases the credibility of the results achieved by the candidates.

The correlation between the written exam results and the results of the practical exam (average since 2011: 0.56; range: 0.48-0.65) suggests that the questions offer a degree of practical relevance given that the tasks set in the OSCE are practice orientated. We believe that, in the future, it would be desirable to verify that those graduates passing this entire exam (written and practical) are also providing their patients with appropriate care as competent doctors. This association is evident from similarly standardised, quality-assured examinations abroad [20], [21], [22], [23], [24].

It is important to note that this article addresses the multiple-choice part only, which becomes a complete examination [2] when supplemented by the previously published clinical skills part [3]. The publication on the clinical skills part of the examination [3] describes how the consistent implementation of the principles of action research contributes to the successful ongoing development of the practical exam. Furthermore, it describes how the centrally coordinated, collaborative iterative process involving experts from all faculties contributes significantly to the quality of the final clinical skills examination (FLE CS) [3]. A qualitative study among participating examination experts and deans of studies also revealed positive implications with respect to the overall examination, such as the intensified and positive collaboration between the faculties and the increased introduction of practical courses [24].

The differences in the pass rates of the graduates of Swiss faculties versus candidates with foreign medical qualifications not recognised in Switzerland can also be understood as supporting the validity of this exam. Such differences in national examinations have also been identified in other countries [25]. One reason could be that the training provided in Switzerland is geared towards the national learning objectives [1], [<http://www.profilesmed.ch/>] such that candidates from Switzerland are better prepared for the examination. Another explanation may be that most candidates with unrecognised foreign medical qualifications have pursued their further training in one discipline and are no longer so familiar with the general exam content. Appreciation for the work of those involved and the central organisation of exam development are two of the major insights gained in recent years that are essential to the long-term success of the process. The appreciation of those involved in

medical teaching is a key aspect, as the clinical and scientific activities are frequently given a higher weighting when they are more conducive to individual careers. It can therefore prove difficult to find suitable individuals for teaching positions and assessment tasks, in particular. This appears to be a common challenge, as (inter)national surveys have shown [26], [27]. According to our experience, by holding national workshops in a seminar hotel to develop the questions and recognising these efforts as a teaching activity – rather than several smaller workshops or working individually at home – participants gain greater appreciation and a strategy is ultimately found that delivers better quality and is more economical [16].

The strengths of this report lie in the fact that it is based on the many years of practical experience gained by the various participants in the development, implementation and evaluation of this national examination, which is currently held in two national languages at five locations. The exam results have proved thoroughly reliable in the last eleven years, which is very much due of course also to the large number of exam questions.

One weakness is that this report is based on experience that has not been subjected specifically to scientific study. However, the insights discussed are based on a continuous optimisation process that is thoroughly documented, traceable, and comprehensible, and has been summarised by the authors for this report.

Looking ahead, it can be said that the development of this examination will steadily continue. Since 2021, the examination has been subject to the new “PROFILES” set of learning objectives [<http://www.profilesmed.ch/>]. The blueprint is now based on this framework, accordingly. The structure of the blueprint, in particular, has therefore been simplified. There are fewer dimensions, for instance, as the previous dimensions of “age” and “gender” and part of the old “medical action” dimension are already considered under the “situations as starting points” of PROFILES, and could be deleted or shortened accordingly. On the other hand, the differentiation under “setting” (practice, hospital, residential care, other) is now greater than in the SCLO (inpatient, outpatient), as the precise setting of the described case is often critical to eliciting the correct answer. Furthermore, the previous dimension of “problems as starting points” has become “situations as starting points” in PROFILES, with physiological situations also listed (e.g. “process and basic care of pregnancy”, “well-baby and well-child visit”), and disciplines such as “emergency” and “palliative care” occupying a more prominent position. It has also been decided that, as of 2022, the written examination will be implemented digitally at all sites using tablet computers. Consequently, new question types can be used in the future that will permit in large part the automated analysis of free texts based on long menus [28]; videos can also be included, moreover.

5. Conclusion

Common guidelines, joint workshops, quality assurance measures with ongoing optimisation of all processes, and appreciation for all involved, are essential to carrying out such an examination at a high-quality level in the long term.

Acknowledgements

A national examination can only be achieved with the support of countless dedicated individuals. Our special thanks go to the participants of the workshops and review boards, and those responsible at the sites. We also thank the board of examiners and the Federal Office of Public Health for their valuable input and support.

Competing interests

The authors declare that they have no competing interests.

References

- Bürgi H, Rindlisbacher B, Bader C, Bloch R, Bosman F, Gasser C, Gerke W, Humair JP, Im Hof V, Kaiser H, Lefebvre D, Schläppi P, Sottas B, Spinaz GA, Stuck AE. Swiss Catalogue of Learning Objectives for Undergraduate Medical Training (SCLLO). Genf: Joint Conference of Swiss Medical Faculties (SMIFK); 2008.
- Guttormsen S, Beyeler C, Bonvin R, Feller S, Schirlo C, Schnabel K, Schurten T, Berendonk C. The new licencing examination for human medicine: from concept to implementation. *Swiss Med Wkly*. 2013;143:w13897. DOI: 10.4414/smw.2013.13897
- Berendonk C, Schirlo C, Balestra G, Bonvin R, Feller S, Huber P, Jünger E, Monti M, Schnabel K, Beyeler C, Guttormsen S, Huwendiek S. The new final Clinical Skills examination in human medicine in Switzerland: Essential steps of exam development, implementation and evaluation, and central insights from the perspective of the national Working Group. *GMS Z Med Ausbild*. 2015;32(4):Doc40. DOI: 10.3205/zma000982
- Swanson DB, Roberts TE. Trends in national licensing examinations in medicine. *Med Educ*. 2016;50(1):101-114. DOI: 10.1111/medu.12810
- Archer J, Lynn N, Coombes L, Roberts M, Gale T, Price T, de Bere SR. The impact of large scale licensing examinations in highly developed countries: a systematic review. *BMC Med Educ*. 2016;16(1):212. DOI: 10.1186/s12909-016-0729-7
- Rimmer A. GMC will develop single exam for all medical graduates wishing to practise in UK. *BMJ*. 2014;349:g5896. DOI: 10.1136/bmj.g5896
- Bundesministerium für Bildung und Forschung. Masterplan Medizinstudium 2020. Berlin: Bundesministerium für Bildung und Forschung; 2017. Zugänglich unter/available from: <https://www.bmbf.de/de/masterplan-medizinstudium-2020-4024.html>
- Gierl MJ, Lai H, Turner SR. Using automatic item generation to create multiple-choice test items. *Med Educ*. 2012;46(8):757-765. DOI: 10.1111/j.1365-2923.2012.04289.x
- Edwards D, Wilkinson D, Canny BJ, Pearce J, Coates H. Developing outcomes assessments for collaborative, cross-institutional benchmarking: Progress of the Australian Medical Assessment Collaboration. *Med Teach*. 2014;36(2):139-147. DOI: 10.3109/0142159X.2013.849798
- Gierl MJ, Lai H. Evaluating the quality of medical multiple-choice items created with automated processes. *Med Educ*. 2013;47(7):726-733. DOI: 10.1111/medu.12202
- Bennett RE. Technology for large-scale assessment. In: Peterson P, Baker E, McGaw B, editors. *International Encyclopaedia of Education*. 3rd ed. Oxford: Elsevier; 2010. p.48-55. DOI: 10.1016/B978-0-08-044894-7.00701-6
- Bennett RE. The changing nature of educational assessment. *Rev Res Educ*. 2015;39:370-407. DOI: 10.3102/0091732X14554179
- Seo DG. Overview and current management of computerized adaptive testing in licensing/certification examinations. *J Educ Eval Health Prof*. 2017;14:17. DOI: 10.3352/jeehp.2017.14.17
- Krebs R. Prüfen mit Multiple Choice. *Kompetent planen, entwickeln, durchführen und auswerten*. Bern: Hofgrefe; 2019. DOI: 10.1024/85092-000
- Bundesamt für Gesundheit. Eidgenössische Prüfung in Humanmedizin. Bern: Bundesamt für Gesundheit; 2019. Zugänglich unter/available from: <https://www.bag.admin.ch/bag/de/home/berufe-im-gesundheitswesen/medizinberufe/eidgenoesische-pruefungen-universitaerer-medizinalberufe/eidgenoesische-pruefung-in-humanmedizin.html>
- Meng-Hentschel J, Delmas C, Zurbuchen B, Wagner F, Schurter T. Erstellung von MC-Fragen für Prüfungen in der Medizin: individuell oder im integrierten Workshop. In: *Gemeinsame Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA) und des Arbeitskreises zur Weiterentwicklung der Lehre in der Zahnmedizin (AKWLZ)*. Münster, 20.-23.09.2017. Düsseldorf: German Medical Science, GMS Publishing House; 2017. Doc353. DOI: 10.3205/17gma353
- Krebs R. The Swiss way to score multiple true-false items: theoretical and empirical evidence. In: Scherpbier AJ, van der Vleuten CM, Rethans JJ, van der Steeg AF, editors. *Advances in Medical Education. Proceedings of the 7th Ottawa International Conference on Medical Education and Assessment*. Dordrecht: Kluwer Academic Publishers; 1996. p.158-161. DOI: 10.1007/978-94-011-4886-3_46
- Angoff WH. Scales, norms and equivalent scores. In: Thorndike RI, editor. *Educational Measurement*. 2nd ed. Washington DC: American Council on Education; 1971. p.508-600.
- Hofstee KW. The case for compromise in educational selection and grading. In: Anderson SB, Helmick JS, editors. *On Educational Testing*. San Francisco: Jossey-Bass; 1983. p.109-127.
- Livingston SA, Zieky MJ. *Passing Scores, A Manual for setting standards of performance on educational and occupational test*. Princeton, NJ: Educational Testing Service; 1982.
- Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: The Danish Institute for Educational Research; 1960.
- Norcini JJ, Boulet JR, Opalek A, Dauphinee WD. The relationship between licensing examination performance and the outcomes of care by international medical school graduates. *Acad Med*. 2014;89(8):1157-1162. DOI: 10.1097/ACM.0000000000000310
- Tamblyn R, Abrahamowicz M, Dauphinee WD, Hanley JA, Norcini J, Girard N, Grand'Maison P, Brailovsky C. Association between licensure examination scores and practice in primary care. *JAMA*. 2002;288(23):3019-3026. DOI: 10.1001/jama.288.23.3019

24. Huwendiek S, Jung D, Schirlo C, Huber P, Balestra G, Guttormsen S, Berendonk C. The introduction of a standardised national licensing exam as a driver of change in medical education: A qualitative study from Switzerland. *Med Teach*. 2020;42(10):1163-1170. DOI: 10.1080/0142159X.2020.1798911
25. Boulet JR, Swanson DB, Cooper RA, Norcini JJ, McKinley DW. A comparison of the characteristics and examination performances of U.S. and non-U.S. citizen international medical graduates who sought Educational Commission for Foreign Medical Graduates certification: 1995-2004. *Acad Med*. 2006;81(10 Suppl):S116-119. DOI: 10.1097/00001888-200610001-00029
26. Huwendiek S, Mennin S, Dern P, Ben-David MF, Van Der Vleuten C, Tönshoff B, Nikendei C. Expertise, needs and challenges of medical educators: Results of an international web survey. *Med Teach*. 2010;32(11):912-918. DOI: 10.3109/0142159X.2010.497822
27. Huwendiek S, Hahn EG, Tönshoff B, Nikendei C. Challenges for medical educators: results of a survey among members of the German Association for Medical Education. *GMS Z Med Ausbild*. 2013;30(3):Doc38. DOI: 10.3205/zma000881
28. Huwendiek S, Reichert F, Duncker C, de Leng BA, van der Vleuten CP, Muijtjens AM, Bosse HM, Haag M, Hoffmann GF, Tönshoff B, Dolmans D. Electronic assessment of clinical reasoning in clerkships: A mixed-methods comparison of long-menu key-feature problems with context-rich single best answer questions. *Med Teach*. 2017;39(5):476-485. DOI: 10.1080/0142159X.2017.1297525

Corresponding author:

Tina Schurter
University of Bern, Institute for Medical Education,
Mittelstr. 43, CH-3012 Bern, Switzerland
tina.schurter@iml.unibe.ch

Please cite as

Schurter T, Escher M, Gachoud D, Bednarski P, Hug B, Kropf R, Meng-Hentschel J, König B, Beyeler C, Guttormsen S, Huwendiek S. *Essential steps in the development, implementation, evaluation and quality assurance of the written part of the Swiss federal licensing examination for human medicine*. *GMS J Med Educ*. 2022;39(4):Doc43. DOI: 10.3205/zma001564, URN: urn:nbn:de:0183-zma0015649

This article is freely available from

<https://doi.org/10.3205/zma001564>

Received: 2021-09-27

Revised: 2022-06-07

Accepted: 2022-07-05

Published: 2022-09-15

Copyright

©2022 Schurter et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 License. See license information at <http://creativecommons.org/licenses/by/4.0/>.

Zentrale Schritte der Erstellung, Durchführung, Auswertung und Qualitätssicherung des schriftlichen Teils der eidgenössischen Prüfung Humanmedizin (Schweizer Staatsexamen)

Zusammenfassung

Zielsetzung: Dieser Bericht beschreibt die zentralen Schritte der Erstellung, Durchführung, Auswertung und Qualitätssicherung des schriftlichen Teils der eidgenössischen Prüfung Humanmedizin (Schweizer Staatsexamen) sowie die seit deren Einführung 2011 gewonnenen Erkenntnisse.

Methodik: Die eidgenössische Prüfung wird auf Basis vorhandener wissenschaftlicher Evidenz, internationaler Expertise sowie Erfahrungen aus vorhergehenden Prüfungen durch Experten aller fünf Medizinischen Fakultäten der Schweiz mit Unterstützung des Instituts für medizinische Lehre erstellt und gleichzeitig an fünf Standorten durchgeführt. Die Prüfungsorganisatoren dokumentieren und reflektieren jede Prüfungsdurchführung, optimieren die Prozesse kontinuierlich und haben die Ergebnisse in diesem Bericht zusammengefasst.

Ergebnisse: Die zentralen Schritte umfassen: Fragenerstellung, -revision, -übersetzung; Prüfungszusammenstellung und Materialherstellung; Kandidatenvorbereitung; Prüfungsdurchführung und -auswertung. Die Qualitätssicherungsmassnahmen bestehen aus Richtlinienkohärenz bei der Fragenerstellung und Prüfungsdurchführung, Revisionsprozessen, Prüfungszusammenstellung auf Basis des nationalen Blueprints, mehrstufiger Überprüfung der Übersetzungen und des Prüfungsmaterials sowie der statistischen Auswertung der Prüfung und der Kandidatenkommentare.

Bewährt hat sich dabei die enge Zusammenarbeit insbesondere von Vertretern aller beteiligten Fakultäten sowie einer zentralen Koordinationsstelle, die sämtliche Schritte methodisch begleitet und sich um die Auswertung der Prüfung kümmert. Die Endergebnisse der Qualitätssicherungsmassnahmen sind erfolgreiche Prüfungsdurchführungen und reliable Ergebnisse in den bisher elf so durchgeführten Prüfungen. Wichtige Erkenntnisse der vergangenen Jahre sind die Bedeutung der Wertschätzung der Arbeit der Beteiligten und die zentrale Organisation der Prüfungserstellung, um den Prozess langfristig erfolgreich durchzuführen.

Schlussfolgerung: Um eine solche Prüfung langfristig mit hoher Qualität durchzuführen, sind gemeinsame Richtlinien und Workshops, Qualitätssicherungsmassnahmen mit fortlaufender Verbesserung aller Prozesse sowie die Wertschätzung der Beteiligten zentral.

Schlüsselwörter: nationale Schlussprüfung, lizenzierende Prüfung, summatives Assessment, Multiple Choice

Tina Schurter¹
Monica Escher²
David Gachoud³
Piotr Bednarski^{4,5}
Balthasar Hug^{6,7}
Roger Kropf^{6,8}
Juliane
Meng-Hentschel¹
Benjamin König¹
Christine Beyeler¹
Sissel Guttormsen¹
Sören Huwendiek¹

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

2 Universität Genf, Medizinische Fakultät, Genf, Schweiz

3 Universität Lausanne, Medizinische Fakultät, Lausanne, Schweiz

4 Universität Fribourg, Medizinische Fakultät, Fribourg, Schweiz

5 Universität Bern, Medizinische Fakultät, Bern, Schweiz

6 Universität Basel, Medizinische Fakultät, Basel, Schweiz

7 Universität Luzern, Medizinische Fakultät, Luzern, Schweiz

8 Universität Zürich, Medizinische Fakultät, Zürich, Schweiz

1. Einleitung

Aufgrund des am 1. September 2007 in Kraft getretenen Bundesgesetzes vom 23. Juni 2006 über die universitären Medizinalberufe (Medizinalberufegesetz, MedBG) musste die eidgenössische Prüfung Humanmedizin zentral organisiert und neu strukturiert werden. Nach entsprechender Vorbereitungszeit wird daher seit 2011 nach Abschluss des Medizinstudiums auf Masterstufe eine nationale (eidgenössische) Prüfung auf Basis der aktuellen Lernziele durchgeführt (2011-2020: *Swiss Catalogue of Learning Objectives for Undergraduate Medical Training SCLO* [1], seit 2021 *Principal Relevant Objectives and Framework for Integrative Learning and Education in Switzerland PROFILES* [<http://www.profilesmed.ch/>]). Nach erfolgreichem Bestehen sind die Kandidaten* dazu berechtigt, den Arztberuf unter Aufsicht auszuüben und die ärztliche Weiterbildung zu beginnen. Die Prüfung besteht aus zwei fächerübergreifenden Teilen [2]: eine Multiple Choice Prüfung (MC) und eine strukturierte klinisch-praktische Prüfung mit Einsatz von standardisierten Patienten (so genannte Clinical Skills Prüfung (CS)) [3]. Diese löst die bisherigen fachspezifischen Schlussprüfungen ab. Bei der Vorbereitung dieser neuen Prüfung besteht die Herausforderung darin, auf möglichst effiziente Weise mit den aktuell fünf medizinischen Fakultäten aus zwei Sprachregionen eine gemeinsame fächerübergreifende, anwendungsorientierte und nach mehrdimensionalem Blueprint gewichtete Prüfung zu erstellen, welche die Befähigung zur Weiterbildung auf nationaler Ebene sicherstellt, den internationalen Qualitätsstandards entspricht und juristischen Anfechtungen standhält.

Das Interesse an nationalen Abschlussprüfungen ist hoch [4], [5]. Weitere Länder denken über deren Neueinführung nach (Grossbritannien [6] und Norwegen), andere über die Vereinheitlichung auch der praktischen Prüfung bei bereits bestehender gemeinsamer schriftlicher Abschlussprüfung (Deutschland [7]). Publikationen zu den Erfahrungen und Erkenntnissen bei der gemeinsamen Erstellung von fakultätsübergreifenden oder nationalen schriftlichen Prüfungen sind jedoch spärlich [8]. Edwards und Team [9] konnten für den australischen Kontext zeigen, dass insbesondere folgende Faktoren den Erfolg einer Prüfungszusammenarbeit ausmachen: *Committed group* (e.g. *medical schools, stakeholders*), *Funding* (e.g. *development grant*), *Engagement* (*Team meetings to build vision*), *Products* (e.g. *framework, assessment development*), *Ownership* (e.g. *Open to all medical schools*). Im Weiteren liegen Publikationen zu speziellen Aspekten vor wie der automatischen Fragen-Generierung für das nationale Examen in Kanada [8], [10] oder zum Einsatz von computerbasiertem [11], [12] oder adaptivem [13] Testen in solch einem Rahmen.

Insgesamt sind nur wenige Publikationen dazu verfügbar, wie eine schriftliche Prüfung von hoher Qualität zielgerichtet für eine nationale Abschlussprüfung erstellt werden kann. Wird der zusätzliche Aspekt der zweisprachigen Durchführung berücksichtigt, so sind noch weniger Erkenntnisse vorhanden.

Der vorliegende Bericht knüpft an bisherige Publikationen zum Schweizer Staatsexamen [2], [3] an. Er stellt dabei die zentralen Schritte, Qualitätssicherungsmaßnahmen und Erkenntnisse aus Sicht der Prüfungsorganisatoren des schriftlichen Teils des Schweizer Staatsexamens nach inzwischen elf Prüfungszyklen vor und bietet Interessierten eine Vergleichsmöglichkeit und Anhaltspunkte für eigene ähnliche Projekte.

2. Methoden

Setting

Die MC-Prüfung umfasst zwei Teilprüfungen mit je 150 MC-Fragen, für die je 4.5 Stunden zur Verfügung stehen, und wird an allen fünf medizinischen Fakultäten gleichzeitig in enger Zusammenarbeit durchgeführt. Bezüglich übergreifender Informationen zu dieser Prüfung verweisen wir auf den bestehenden Übersichtsartikel [2]. Nebst inhaltlichen Experten aus den Kliniken sind Vertreter der Hausarztmedizin, methodische Experten (Expertise in Fragenerstellung, -revision und Auswertung), Informatiker, Sachbearbeiter, professionelle Übersetzer, Vertreter des Bundesamtes für Gesundheit (BAG) sowie die eidgenössische Prüfungskommission am Prüfungsprozess beteiligt. Die Prüfungskommission als wichtigstes Entscheidungsgremium besteht aus den Vizedekanen für Lehre der medizinischen Fakultäten sowie Vertretern der Hausarztmedizin, der Weiterbildung und des BAG. Die methodischen Experten konstituieren die nationale Arbeitsgruppe MC mit Vertretern der Fakultäten und des Instituts für Medizinische Lehre (IML). Abgesehen von der Rekrutierung der Fragenautoren und der Durchführung der Prüfung in den Fakultäten erfolgt die Koordination durch das IML.

Erstellung dieses Projektberichts

Dieser Projektbericht beruht auf den Erfahrungen aus bisher elf schriftlichen eidgenössischen Prüfungen seit 2011. Nach jedem Prüfungsdurchgang werden die aktuelle Prüfung inklusive Qualitätssicherungen und Qualitätsindizes reflektiert und dokumentiert, um Bewährtes fortzusetzen und nach Optimierungsmöglichkeiten zu suchen. In einem iterativen Prozess werden die zentralen Schritte der Erstellung, Durchführung, Auswertung und Qualitätssicherung sowie die Erkenntnisse durch die Prüfungsorganisatoren zusammengetragen.

3. Ergebnisse

Der Prüfungsprozess wird nachfolgend detailliert erklärt. Abbildung 1 zeigt dabei die wichtigsten Schritte des Prozesses, Tabelle 1 fasst die Qualitätssicherung und die Erkenntnisse zusammen.



Abbildung 1: Zentrale Schritte des Prüfungsprozesses

3.1. Fragenstellung und Revision

Rund 60 von den Fakultäten benannte Fachexperten erstellen Fragenentwürfe, die am ersten Tag eines zweitägigen Workshops in Autorengruppen (Peer-Review) diskutiert und gegebenenfalls revidiert werden. Am zweiten Workshoptag werden die neuen Fragen in Revisionsgruppen, bestehend aus mindestens einem Vertreter jeder Fakultät und Vertreter unterschiedlicher medizinischer Fachrichtungen, ein zweites Mal begutachtet und finalisiert. In den gleichen Revisionsgruppen werden zudem alle alten Fragen auf Aktualität überprüft und angepasst, die fünf Jahre zuvor erstellt wurden oder die in der vergangenen Prüfung durch statistische Abweichungen respektive Kandidatenkommentare aufgefallen waren. Nach der zentral durchgeführten formalen und sprachlichen Überprüfung der neuen Fragen durch methodische Experten des IML werden die Fragen in einem nationalen Review-Board (bestehend aus Medizinerinnen verschiedener Fachrichtungen aus allen Fakultäten, die die Richtlinien der Fragenerstellung dieser Prüfung gut kennen und auch selber als Autoren Fragen erstellen) auf Stufengerechtigkeit (angemessener Schwierigkeitsgrad) und klinische Relevanz überprüft, bevor sie in der Datenbank für die Zusammenstellung der nächsten Prüfung freigegeben werden. Die Fragenverwaltung findet innerhalb einer am IML entwickelten Software statt [<https://www.iml.unibe.ch/themen/uebersichten/projekte>].

Qualitätssicherung

Die Anforderungen an die Fragenautoren umfassen die inhaltlichen Grundlagen (Lernziele und darauf basierender inhaltlicher Blueprint), Informationen über die Fragetypen,

die vorgegebene Standardstruktur für Fallvignetten sowie Formulierungsprinzipien gemäss Krebs [14]. Bevorzugt werden Fragen mit Fallvignetten und damit anwendungsorientierte Fragen unter anderem zu diagnostischen oder therapeutischen Entscheiden erstellt.

In den öffentlichen Vorgaben der Medizinalberufekommission (MEBEKO) ist zum Anwendungsbezug der Fragen Folgendes definiert [15]: „Mit der Clinical Knowledge Prüfung (schriftliche Prüfung) wird fächerübergreifend anwendungsorientiertes Wissen zum gesamten Spektrum humanmedizinischer Probleme geprüft“ und „Die Fragen sollen sich wenn immer möglich auf ein konkretes Problem beziehen, das in einer so genannten Fall- oder Problemvignette präsentiert wird.“ In den internen Anleitungen resp. Schulungsunterlagen für die Autoren wird dies mit Beispielen konkret spezifiziert. Jede Frage soll möglichst authentisch sein, also den Arbeitsalltag eines Arztes widerspiegeln. Dazu soll sie sich wenn möglich auf einen konkreten Fall beziehen, in dem eine Patientin oder ein Patient beschrieben wird (Alter, Geschlecht, Setting, Konsultationsgrund, Anamnese, Status/Befunde, ev. Resultate der diagnostischen Untersuchungen, ev. Anfangsbehandlung, Folgebefunde etc.), zu dem eine bestimmte Frage beantwortet werden muss, z.B. „Welches ist die wahrscheinlichste Diagnose?“, „Was ist therapeutisch am sinnvollsten?“. Zudem sollte jede Frage auch von klinischer Relevanz sein.

Zu den Fragetypen ist in den MEBEKO-Vorgaben definiert [15]: „Auswahl der einzig richtigen oder besten aus 3-5 angebotenen Wahlantworten (Typ A, positiv oder negativ formuliert) und vierfache Entscheidung richtig/falsch (Typ Kprim).“

Welche Fragen erstellt werden, wird anhand der Lernziele und des Blueprints (siehe Tabelle 2) festgelegt. Im Blue-

Tabelle 1: Zusammenfassung der Qualitätssicherung und Erkenntnisse

Schritte	Qualitätssicherung	Erkenntnisse
Fragenerstellung und Revision	<ul style="list-style-type: none"> • Gemeinsame Richtlinien zur Fragenerstellung • Gruppenarbeit mit individueller Vorbereitung • Einbezug aller Fakultäten in allen Revisionschritten inhaltlich (u.a. bezüglich der Authentizität und klinischen Relevanz der Fragen) & formal • Befragung der Workshopteilnehmer 	Ein gemeinsamer 2-tägiger Workshop führt zu deutlich grösserem Output als viele halbtägige Workshops in kleinen (lokalen) Gruppen und wird von den Experten als wertschätzend und motivierend empfunden.
Übersetzung	<ul style="list-style-type: none"> • Mehrstufige Überprüfung der Übersetzung • Ausschluss von fehlerhaft übersetzten Fragen aus der Bewertung 	Eine zusätzliche Überprüfung der Übersetzung durch eine medizinische Fachperson in ihrer Muttersprache verbessert die Übersetzungen.
Prüfungszusammenstellung	<ul style="list-style-type: none"> • Einhaltung des Blueprints • Auswahl zuvor genutzter Fragen nach guter Performance • Nutzung der Ergebnisse der Befragung der Kandidaten nach der Prüfung 	Verschlagwortung der Fragen ist essentiell für eine durchsuchbare Datenbank. Die Auswertungs-Daten sind ebenfalls such- und filterbar.
Erstellung Prüfungsmaterial	<ul style="list-style-type: none"> • Mehrstufiges Lektorat der Prüfungshefte 	Ein Lektorat ist nötig, da dabei gelegentlich noch Auffälligkeiten festgestellt werden.
Kandidatenvorbereitung	<ul style="list-style-type: none"> • Jährliche Überprüfung und Anpassung der Informationen • Erstellung und Pflege eines Self Assessment Tools online, zur formellen Vorbereitung auf Niveau und Formate • Informationsveranstaltungen für die Kandidaten in den Fakultäten • Analyse der Einsprachen gegen das Prüfungsergebnis 	Nicht alle Kandidaten informieren sich aktiv über die Prüfung, wenn das Informationsmaterial nur online zur Verfügung steht.
Prüfungsdurchführung	<ul style="list-style-type: none"> • Gemeinsame Richtlinien zur Prüfungsdurchführung • Nutzung der Ergebnisse der Befragung der Kandidaten nach der Prüfung 	Einigen Kandidaten fällt es schwer, am Prüfungstag Informationen oder Anweisungen aufzunehmen.
Auswertung	<ul style="list-style-type: none"> • Befragung der Kandidaten nach der Prüfung und u.a. Analyse der Kommentare zu Mängeln einzelner Prüfungsfragen • Vieraugenprinzip der wichtigsten Schritte • Ausschluss von inhaltlich oder formal mangelhaften Prüfungsfragen aus der Bewertung und Rückmeldung der Mängel an Autoren und Review-Board • Item-Statistik <i>Discrimination</i> • Item-Statistik <i>Difficulty</i> • Reliabilität • Bestehensgrenzenfestlegung 	Dank der Rückmeldung vergangener Mängel konnte die Fragenerstellung laufend optimiert werden, sodass über die Jahre immer weniger Fragen aufgrund formaler oder inhaltlicher Mängel aus der Bewertung ausgeschlossen werden mussten.

print ist auch die prozentuale inhaltliche Verteilung der Fragen festgelegt. Alle Autoren sind in die Lehre sowie die Erstellung von Fragen für fakultäre Prüfungen involviert. Im Rahmen des Workshops werden sie durch Informationsmaterial, das sie zur Erstellung ihrer Fragenentwürfe vor dem Workshop erhalten, mit den Standards der eidgenössischen Prüfung vertraut gemacht. Der Einbezug aller Fakultäten in allen Revisionschritten (inklusive in Aus-

nahmefällen Streichung von Fragen) ermöglicht die Erstellung einer national gültigen und von allen Ausbildungsstellen akzeptierten Prüfung.

Erkenntnisse

Die Befragung der Teilnehmenden der Workshops zeigt, dass der Rahmen der Fragenerstellung seit 2017

Tabelle 2: Aktueller Blueprint (BP) (entsprechend PROFILES, gültig ab 2021) für den schriftlichen Teil der Eidgenössischen Prüfung Humanmedizin – Hauptdimensionen 1 und 2 (englische Version)

Dimension 1: Situations as Starting Point (SSP)	Dimension 2: Medical Task		
	1. Understanding mechanisms of disease	2. Assessment / Diagnosis	3. Management
1. General complaints and symptoms (SSP 1-16)			
2.1 Head and neck (SSP 17-39)			
2.2 Chest (SSP 40-51)			
2.3 Abdomen (SSP 52-61)			
2.4 Pelvis, urogenital system (SSP 62-80)			
2.5 Back and extremities (SSP 81-86)			
2.6 Skin (SSP 87-95)			
2.7 Nervous system (SSP 96-105)			
2.8 Injuries and trauma (SSP 106-117)			
2.9 Emotional and behavioural symptoms (SSP 118-130)			
3.1 Findings upon physical examination (SSP 131-149)			
3.2 Findings upon additional examination (SSP 150-174)			
4.1 Situations related to pregnancy and motherhood (SSP 175-183)			
4.2 Situations related to childhood (SSP 184-195)			
4.3 Situations related to old age (SSP 196-202)			
4.4 Emergency symptoms, findings and situations (SSP 203-220)			
4.5 Issues linked with prevention and health promotion (SSP 221-227)			
4.6 Palliative care (SSP 228-232)			
4.7 Psychosocial issues (SSP 233-241)			
4.8 Various health care issues (SSP 242-265)			

(Workshop ohne Alltagsablenkung, Austausch mit Fachkollegen schweizweit) als sehr motivierend empfunden wird [16], im Gegensatz zu den bis 2016 dezentral und in Einzelarbeit erstellten Fragen. Zudem führt das Workshop-Format zu einem erhöhten Output an neuen Fragen mit höherer Qualität (geringerer Revisionsbedarf). Die auf zwei Tage konzentrierte Arbeit wird durch die oben genannte webbasierte, gesicherte Fragendatenbank ermöglicht.

3.2. Übersetzung

Die Fragen werden je nach Autor auf Deutsch oder Französisch erstellt und nach dem Revisionsprozess durch professionelle Übersetzer in die jeweils andere Sprache übersetzt. Die Übersetzung wird sowohl formal und sprachlich durch methodische Experten als auch inhaltlich durch medizinische Fachexperten überprüft.

Qualitätssicherung

Die Qualität der Übersetzungen wird durch die mehrstufige Überprüfung sichergestellt. Nach der Prüfung werden die Kandidatenkommentare und die Antwortungsmuster der Fragen analysiert, um Übersetzungsfehler oder

fachliche Unklarheiten aufzudecken. Fragen mit einem Übersetzungsfehler werden wie andere Fragen mit möglichen formalen Mängeln aus der Bewertung ausgeschlossen.

Erkenntnisse

Teilweise wurden in den Kandidatenkommentaren einzelne Übersetzungsprobleme genannt. Deshalb werden inzwischen alle Fragen, zusätzlich zum bestehenden Kontrollprozess, einer medizinischen Fachperson in ihrer Muttersprache vorgelegt.

Die Anpassung und Einführung zusätzlicher Kontrollschritte in der Übersetzungskontrolle hat dazu geführt, dass weniger Fragen aufgrund mangelhafter Übersetzungen eliminiert werden mussten.

3.3. Prüfungszusammenstellung

Pro Prüfung werden 300 Fragen aus der Fragendatenbank gemäss national geltendem Blueprint (erste zwei Dimensionen siehe Tabelle 2) zusammengestellt, der auf Basis der Lernziele entwickelt wurde [1], [<http://www.profilesmed.ch/>]. Der aktuelle Blueprint besteht aus drei Hauptdimensionen:

- Dimension 1: *Situations as starting points*,
- Dimension 2: *Medical tasks* und
- Dimension 3: *General objectives* (Fokus auf *medical expert*). *Type of condition* (*acute, subacute, chronic*) und *setting* (*ambulatory practice, hospital, nursing home for elderly people, other*) stellen weitere Dimensionen des Blueprints dar.

Die prozentualen Sollwerte werden nicht veröffentlicht, daher in der Tabelle auch nicht angegeben, um damit ein strategisches Auslassen einzelner Wissensgebiete beim Lernen zu verhindern. Es wird dabei ein Anteil von mindestens 20% bewährter Fragen früherer Prüfungen eingesetzt. Durch die klaren Vorgaben durch das Prüfungsreglement, inhaltlich anhand des Blueprints und der Vorgaben bezüglich der verwendeten Fragetypen wird die Kongruenz von Lernzielen und Prüfungsfragen gewährleistet.

Qualitätssicherung

Die repräsentative Verteilung der Inhalte wird durch den Blueprint und die Befragung der Kandidaten im Anschluss an die Prüfung anhand eines Fragebogens mit offenen und geschlossenen Fragen sichergestellt. Die Einhaltung des Blueprints wird dank der Verschlagwortung der Fragen in der Fragendatenbank unterstützt.

Erkenntnisse

Die Auswertung des Fragebogens zeigt, dass vielen Kandidaten eine möglichst breite Abdeckung des Lernstoffs wichtig ist (Kommentare zu offenen Fragen des Fragebogens), und dass der Inhalt der Fragen mit den Lernzielen übereinstimmt (mit einem Median von 4 auf der Skala von 1 (überhaupt nicht einverstanden) bis 5 (absolut einverstanden), geschlossene Frage des Fragebogens). Die Einhaltung des Blueprints und die breite Abdeckung ist nur dank der detaillierten Verschlagwortung der Fragen möglich.

3.4. Erstellung Prüfungsmaterial

Die Prüfungsheftvorlagen werden aus der Fragendatenbank heraus erzeugt und gedruckt, bevor sie an die Fakultäten verschickt werden. Die Themengebiete der Fragen werden gleichmässig auf beide Prüfungshefte verteilt, auch die Textlänge beider Prüfungshefte ist gleich.

Qualitätssicherung

Die Qualität der Prüfungshefte wird mittels eines mehrstufigen Lektorats durch inhaltliche und methodische Überprüfung (Form und Sprache) durch Experten sichergestellt.

Erkenntnisse

Trotz vorangehendem Revisionsprozess bei der Fragenstellung und Übersetzung werden beim Lektorat der

Druckvorlagen gelegentlich noch Auffälligkeiten festgestellt. Diese werden zur inhaltlichen Überprüfung durch Ärzte gegengelesen und berichtet.

3.5. Kandidatenvorbereitung

Die Kandidaten informieren sich auf der Homepage des Bundesamtes für Gesundheit sowie in fakultären Informationsveranstaltungen über den Prüfungsablauf [15] und erhalten von den Standortverantwortlichen schweizweit einheitliche Unterlagen. Zudem werden rund 300 repräsentative Fragen online in einem Self-Assessment-Tool zur Verfügung gestellt [<https://www.iml.unibe.ch/angebote/assessment/pruefungsdienstleistungen/self-assessment>].

Qualitätssicherung

Die vorbereitenden Informationen werden jährlich überprüft und nötigenfalls angepasst.

Erkenntnisse

Rückmeldungen der Kandidaten im Fragebogen sowie Argumente in Einsprachen gegen das Prüfungsergebnis zeigten auf, dass nicht alle Kandidaten die vor der Prüfung zur Verfügung stehenden Informationen aktiv einholen. Deshalb werden diese den Kandidaten im Vorfeld zusätzlich persönlich zugeschickt.

3.6. Prüfungsdurchführung

Die Prüfung wird an zwei Tagen zur selben Zeit (4,5 Std.) an den fünf Fakultäten durchgeführt. Zur Prüfung treten nebst den erfolgreichen Absolventen der Masterstudiengänge der Schweiz auch Kandidaten mit nicht anerkannten ausländischen Arzt diplomen an, welche ausserhalb der EU/EFTA ausgestellt wurden.

Qualitätssicherung

Gemeinsame Richtlinien zur Prüfungsdurchführung regeln beispielsweise die Gegebenheiten der Prüfungsräume, erlaubten Hilfsmittel oder während der Prüfung vorzulesenden Instruktionen.

Erkenntnisse

Einigen Kandidaten fällt es schwer, am Prüfungstag Informationen oder Anweisungen zum Prüfungsablauf aufzunehmen. Deshalb werden ihnen sämtliche Informationen zusätzlich im Vorfeld zur Verfügung gestellt. Die Vorschläge der Kandidaten im Fragebogen helfen, die Durchführung laufend zu verbessern.

3.7. Auswertung und Ergebnisse

Für jede verwendete Prüfungsfrage wird eine Item-Analyse (Beurteilung der Messeigenschaften der Items) durchgeführt. Primär beurteilt werden Schwierigkeit und Trenn-

schärfe; diese bleiben im Fragen-Pool sichtbar. Bei Wiederverwendung derselben Frage kann diese somit über verschiedene Prüfungs-Jahrgänge miteinander verglichen werden. Als Trennschärfe-Index r eines Items wird mindestens 0,2 angestrebt und als Schwierigkeitsgrad ein P von 50-90%. Die multiplen Falsch/Richtig-Fragen vom Typ Kprim werden nach der am IML entwickelten Halbpunkt-Methode ausgewertet [17].

Die Prüfungsergebnisse und die Kommentare der Kandidaten zu mangelhaft erscheinenden Fragen werden zunächst durch methodische Experten analysiert. Auffällige Fragen werden mit mehreren klinischen Fachexperten diskutiert und, falls sich dabei formale oder inhaltliche Mängel zeigen, von der Bewertung ausgeschlossen. Zum Beispiel lässt sich bei einer Frage mit zu hoher Schwierigkeit ein „Facharzt“- Niveau feststellen, oder bei zu geringem r ein zuvor übersehener formaler Fehler erkennen. Die Bestehensgrenze beruht auf zwei international bekannten Standardsetting-Methoden (nach Angoff und Hofstee [18], [19], [20]) sowie einer Analyse auf Grundlage des Rasch-Modells [21]. Das Rasch-Modell ist das einfachste Modell der Item-Response Theorie (IRT), bei welchem die Item-Schwierigkeiten und die Fähigkeiten der Kandidierenden geschätzt werden, um damit die Prüfungsergebnisse zu erklären. Das Modell erlaubt es, durch den Einsatz von Items, deren Schwierigkeit bereits über vergangene Prüfungen beobachtet und deshalb bekannt sind (sogenannte Ankeritems), die Bestehensanforderung über die Jahre konstant zu halten, auch wenn die Prüfungen in den einzelnen Jahren unterschiedlich schwierig ausfallen. Einfach gesagt ermöglicht ein Vergleich zwischen Ankeritems und dem Rest der Prüfung eine Einschätzung der Prüfungsschwierigkeit im Vergleich zu den Vorjahren und die Bestehensgrenze kann somit entsprechend angepasst werden. Auf Basis der Standardsetting-Methoden und dem Raschmodell wird dann die Bestehensgrenze durch die Prüfungskommission festgelegt.

Qualitätssicherung

Die endgültige Entscheidung über die Bestehensgrenze wird durch die nationale Prüfungskommission gefällt. Die bei der Auswertung entdeckten formalen oder inhaltlichen Mängel an Prüfungsfragen werden den Autoren und dem Review-Board gemeldet, damit die nächste Fragenproduktion entsprechen optimiert werden kann. Fragen mit schlechten Messwerten der Schwierigkeit und Trennschärfe und inhaltlichen Mängeln werden nicht wiederverwendet. Alle wichtigen Auswertungsschritte werden im Vieraugenprinzip durchgeführt: Eine zweite Person kontrolliert die wichtigsten Ergebnisse anhand einer Checkliste, bevor sie veröffentlicht werden. Beispielsweise: Wurde auf den Resultatelisten, in den Grafiken im Analysebericht und auf den Briefen an die Kandidaten wirklich die letztendlich durch die Prüfungskommission festgelegte und im Sitzungsprotokoll festgehaltene Bestehensgrenze angewendet? Stimmen die Antworten auf den Antwortbögen der nicht bestehenden Kandidaten tatsächlich mit

den elektronisch erfassten und für die Resultateberechnung berücksichtigten Antworten überein? Die Auswertungen bilden einen zentralen letzten Schritt der Qualitätssicherung des gesamten Prüfungsprozesses, weitere Details sind unter [14], Kapitel 4, dargestellt.

Erkenntnisse

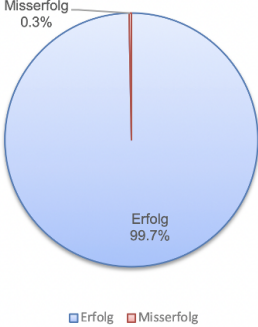
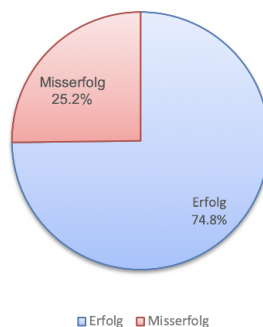
Dank kontinuierlich optimierten Massnahmen zur Qualitätssicherung im gesamten Prüfungsprozess, einschliesslich der Rückmeldung der entdeckten Mängel in Prüfungsfragen an die Autoren und das Review-Board, mussten über die Jahre immer weniger Fragen aufgrund formaler oder inhaltlicher Mängel aus der Bewertung ausgeschlossen werden (rund 40 in der ersten Prüfung (13.3%), zuletzt noch rund 15 (5%)). Die Ergebnisse der bisherigen elf Prüfungsdurchführungen zeigten durchwegs reliable Prüfungsergebnisse (Cronbach Alpha im Mittel 0.90, Range 0.87-0.91).

Abbildung 2 zeigt exemplarisch die Ergebnisse der Prüfung 2020. Die Erfolgsraten der Kandidaten, die ihr Studium an einer Schweizer Fakultät abgeschlossen haben (fakultäre Kandidaten genannt), waren mit 99.5% konstant hoch, währenddem die Erfolgsraten der Kandidaten mit nicht anerkanntem ausländischen Arzt Diplom deutlich tiefer lagen. Dies könnte so interpretiert werden, dass Schweizer Kandidierende auch auf Basis des Schweizer Lernzielkatalogs unterrichtet wurden und somit bei diesen eine gute Abstimmung von Unterricht und Prüfung vorliegt, bei den ausländischen Kandidierenden jedoch nicht. Weitere Gründe könnten in den sprachlichen Schwierigkeiten der ausländischen Kandidaten liegen oder darin, dass deren Abschluss oft mehrere Jahre zurückliegt und das Wissen daher nicht mehr gleich präsent ist.

4. Diskussion

Wichtige Qualitätssicherungsmaßnahmen umfassen unter anderem die gemeinsamen Richtlinien zur Fragenstellung und Prüfungsdurchführung, die Fragenstellung in nationalen Workshops, den mehrstufigen Revisionsprozess unter Einbezug aller Fakultäten, die Prüfungszusammenstellung auf Basis des nationalen Blueprints sowie die mehrstufige Überprüfung der Übersetzungen und des Prüfungsmaterials. Die Prüfungsanalyse inklusive Analyse der Kandidatenkommentare ist ein abschliessender wichtiger Qualitätssicherungsschritt des gesamten Prüfungsprozesses.

Während über einige dieser Prozesse einschliesslich Erfolgsfaktoren auch von Edwards und Kollegen [9] berichtet wurde, können wir zusätzlich über Aspekte der Fragenstellung in nationalen Workshops, der Zweisprachigkeit und der Auswertung berichten. Hinzu kommt, dass der Einbezug aller Fakultäten bei der Erstellung des Blueprints sowie in allen Fragenstellungs-, Revisions-, und Durchführungsschritten zu einer hohen Akzeptanz bei den Fa-

Ergebnisse 2020	Fakultäre Kandidaten**	Kandidaten mit nicht anerkenbarem ausländischen Diplom***
Total Kandidaten	1105	103
Mittlere Leistung (Anteil erreichter Punkte)	76.6%	60.6%
Standardabweichung	6.9%	11.5%
Anteile Erfolg/Misserfolg	 <p>Misserfolg 0.3% Erfolg 99.7%</p>	 <p>Misserfolg 25.2% Erfolg 74.8%</p>

* Die Resultate 2020 sind vergleichbar für alle Resultate seit 2011.

** Fakultäre Kandidaten: Kandidaten, die ihr Studium an einer Schweizer Fakultät abgeschlossen haben und zum ersten Mal zur Prüfung antreten.

*** Kandidaten mit nicht anerkenbarem ausländischen Diplom: Kandidaten, die ihr Studium ausserhalb der Schweiz und der EU/EFTA absolviert haben und zum ersten Mal zur Prüfung antreten.

Abbildung 2: Ergebnisse der Prüfung 2020*

kultäten führt und die Glaubwürdigkeit der Kandidatenergebnisse erhöht.

Die Korrelation der schriftlichen Prüfungsergebnisse mit denen der praktischen Prüfung (Mittelwert seit 2011: 0.56, Range: 0.48-0.65) legt eine gewisse Praxisrelevanz der Fragen nahe, da die Aufgabenstellungen im OSCE ja praxisorientiert sind. Erstrebenswert erscheint uns in Zukunft insbesondere die Überprüfung, dass diejenigen Absolventen, die diese Gesamt-Prüfung (schriftliche und praktische Prüfung) bestehen, danach auch als kompetente Ärzte ihre Patienten zielführend versorgen. Ähnlich standardisierte und qualitätssgesicherte Prüfungen aus dem Ausland legen diesen Zusammenhang nahe [22], [23], [24].

Wichtig ist zu berücksichtigen, dass im vorliegenden Artikel nur über den Multiple Choice-Teil berichtet wird, der durch den zuvor publizierten Clinical Skills-Teil [3] zu einer Gesamt-Prüfung [2] ergänzt wird. In der Publikation zum Clinical Skills Teil der Prüfung [3] wird vorgestellt, dass die konsequente Umsetzung der Prinzipien der Aktionsforschung zur erfolgreichen Weiterentwicklung des praktischen Teils der Prüfung beiträgt. Weiterhin wird vorgestellt, dass der zentral koordinierte, kollaborativ-iterative Prozess mit Einbindung von Fachexperten aus allen Fakultäten wesentlich zur Qualität der EP CS beiträgt [3]. Für die Gesamtprüfung wurden auch positive Implikationen in einer qualitativen Studie unter beteiligten Prüfungsexperten und Studiendekanen gesehen, wie beispielsweise die intensiviertere und positive Zusammenarbeit der Fakultäten und die vermehrte Einführung praktischer Kurse [24].

Die Unterschiede der Bestehensraten der Absolventen der Schweizer Fakultäten gegenüber den Kandidaten mit

in der Schweiz nicht anerkenbaren ausländischen Arzt-diplomen kann auch als Unterstützung der Validität dieser Prüfung verstanden werden. Derartige Unterschiede in nationalen Examen wurden auch in anderen Ländern festgestellt [25]. Eine Ursache könnte darin liegen, dass sich die Ausbildung in der Schweiz an den nationalen Lernzielen [1], [<http://www.profilesmed.ch/>] ausrichtet und daraus für die Kandidaten aus der Schweiz eine bessere Prüfungsvorbereitung resultiert. Als weitere Ursache kommt in Frage, dass die meisten Kandidaten mit nicht anerkenbaren ausländischen Arzt-diplomen ihre Weiterbildung in einem Fachgebiet vorangetrieben haben und mit den allgemeinen Prüfungsinhalten nicht mehr so vertraut sind. Zu den wichtigsten Erkenntnissen der vergangenen Jahre gehört, dass die Wertschätzung der Arbeit aller Beteiligten sowie die zentrale Organisation der Prüfungserstellung essentiell sind, um den Prozess langfristig erfolgreich durchführen zu können. Die Wertschätzung der an der medizinischen Lehre beteiligten Personen stellt einen zentralen Gesichtspunkt dar, da die klinische und wissenschaftliche Arbeit häufig höher gewichtet werden, die der individuellen Karriere förderlicher sind. Dadurch kann es schwierig werden, geeignete Personen für Lehr- und insbesondere Assessmentaufgaben zu finden. Dies scheint eine verbreitete Herausforderung zu sein, wie (inter-)nationale Umfragen zeigen [26], [27]. Gemäss unseren Erfahrungen konnte durch die Fragenerstellung im Rahmen nationaler Workshops in einem Seminarhotel unter Anerkennung der Arbeit als Lehrleistung – anstelle mehrerer kleinerer Workshops beziehungsweise individueller Heimarbeit – eine höhere Wertschätzung der Teilnehmer und im Endeffekt ein

qualitativ besserer und wirtschaftlich günstigerer Weg gefunden werden [16].

Die Stärken des vorliegenden Berichts bestehen darin, dass dieser auf den langjährigen praktischen Erfahrungen der zahlreichen Beteiligten bei der Erstellung, Durchführung und Auswertung dieser nationalen Prüfung beruht, welche aktuell in zwei Landessprachen an fünf Standorten durchgeführt wird. Dabei zeigten sich die letzten elf Jahre durchwegs reliable Prüfungsergebnisse, wobei dies natürlich auch sehr durch die grosse Anzahl an Prüfungsfragen unterstützt wird.

Eine Schwäche liegt darin, dass dieser Bericht auf Erfahrungen basiert, die nicht im Einzelnen gezielt wissenschaftlich untersucht wurden. Allerdings beruhen die hier vorgestellten Erkenntnisse auf einem kontinuierlichen Verbesserungsprozess, der gut dokumentiert, nachvollziehbar und nachvollziehbar ist und von den Autoren für diesen Bericht zusammengefasst wurde.

Ausblickend lässt sich feststellen, dass die Fortentwicklung dieser Prüfung kontinuierlich weitergeht. Seit 2021 ist das neue Lernzielrahmenwerk „PROFILES“ für diese Prüfung gültig [<http://www.profilesmed.ch/>]. Dementsprechend basiert der Blueprint neu darauf. Dabei wurde insbesondere die Struktur des Blueprints vereinfacht. So gibt es zum Beispiel weniger Dimensionen, da die alten Dimensionen „Alter“ und „Geschlecht“ sowie ein Teil der alten Dimension „Ärztliche Handlung“ unter PROFILES bereits in den „Ausgangssituationen“ (original „Situations as Starting Points“) berücksichtigt sind und entsprechend gelöscht oder gekürzt werden konnten. Hingegen wird in der Dimension „Setting“ neu differenzierter unterschieden (Praxis, Krankenhaus, Altersheim, andere) als noch unter SCLO (stationär, ambulant), da das präzise Setting des beschriebenen Falles oft ausschlaggebend für die korrekte Antwort ist. Zudem wurde die alte Dimension Konsultationsgründe/Leitsymptome (original „Problems as Starting Points“) mit PROFILES zu „Ausgangssituationen“ (original „Situations as Starting Points“), wobei nun auch physiologische Situationen vorkommen (z.B. „Vorsorgeuntersuchungen in der Schwangerschaft“ (original „process and basic care of pregnancy“ oder „pädiatrische Vorsorgeuntersuchungen“ (original „well-baby and well-child visit“) und Bereiche wie „Notfall“ (original „Emergency“) oder Palliativmedizin/Palliativpflege (original „Palliative care“) einen wichtigeren Platz einnehmen. Zudem ist entschieden, ab 2022 die schriftliche Prüfung an allen Standorten elektronisch auf Tabletcomputern durchzuführen. Dadurch können in Zukunft einerseits neue Fragetypen verwendet werden, die beispielsweise die Auswertung von Freitexten anhand von so genannten Long Menus [28] zu einem grossen Teil automatisiert ermöglichen, andererseits können neu auch Videos eingesetzt werden.

5. Schlussfolgerung

Um eine solche Prüfung langfristig mit hoher Qualität durchzuführen, sind gemeinsame Richtlinien, gemeinsa-

me Workshops, Qualitätssicherungsmaßnahmen mit fortlaufender Verbesserung aller Prozesse sowie die Wertschätzung der Beteiligten zentral.

Anmerkung

*Für eine einfachere Lesbarkeit wurde die männliche Form gewählt, gemeint sind damit jedoch alle Personen.

Danksagung

Eine nationale Prüfung lässt sich nur durch Einsatz von unzähligen engagierten Personen bewerkstelligen. Unser ausgesprochener Dank geht insbesondere an die Teilnehmer der Workshops und Review-Boards sowie an die Standortverantwortlichen. Im Weiteren danken wir der Prüfungskommission und dem Bundesamt für Gesundheit für die gute Zusammenarbeit und wertvolle Unterstützung.

Interessenkonflikt

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

Literatur

- Bürgi H, Rindlisbacher B, Bader C, Bloch R, Bosman F, Gasser C, Gerke W, Humair JP, Im Hof V, Kaiser H, Lefebvre D, Schläppi P, Sottas B, Spinass GA, Stuck AE. Swiss Catalogue of Learning Objectives for Undergraduate Medical Training (SCLO). Genf: Joint Conference of Swiss Medical Faculties (SMIFK); 2008.
- Guttormsen S, Beyeler C, Bonvin R, Feller S, Schirlo C, Schnabel K, Schurten T, Berendonk C. The new licencing examination for human medicine: from concept to implementation. *Swiss Med Wkly.* 2013;143:w13897. DOI: 10.4414/smw.2013.13897
- Berendonk C, Schirlo C, Balestra G, Bonvin R, Feller S, Huber P, Jünger E, Monti M, Schnabel K, Beyeler C, Guttormsen S, Huwendiek S. The new final Clinical Skills examination in human medicine in Switzerland: Essential steps of exam development, implementation and evaluation, and central insights from the perspective of the national Working Group. *GMS Z Med Ausbild.* 2015;32(4):Doc40. DOI: 10.3205/zma000982
- Swanson DB, Roberts TE. Trends in national licensing examinations in medicine. *Med Educ.* 2016;50(1):101-114. DOI: 10.1111/medu.12810
- Archer J, Lynn N, Coombes L, Roberts M, Gale T, Price T, de Bere SR. The impact of large scale licensing examinations in highly developed countries: a systematic review. *BMC Med Educ.* 2016;16(1):212. DOI: 10.1186/s12909-016-0729-7
- Rimmer A. GMC will develop single exam for all medical graduates wishing to practise in UK. *BMJ.* 2014;349:g5896. DOI: 10.1136/bmj.g5896
- Bundesministerium für Bildung und Forschung. Masterplan Medizinstudium 2020. Berlin: Bundesministerium für Bildung und Forschung; 2017. Zugänglich unter/available from: <https://www.bmbf.de/de/masterplan-medizinstudium-2020-4024.html>
- Gierl MJ, Lai H, Turner SR. Using automatic item generation to create multiple-choice test items. *Med Educ.* 2012;46(8):757-765. DOI: 10.1111/j.1365-2923.2012.04289.x

9. Edwards D, Wilkinson D, Canny BJ, Pearce J, Coates H. Developing outcomes assessments for collaborative, cross-institutional benchmarking: Progress of the Australian Medical Assessment Collaboration. *Med Teach.* 2014;36(2):139-147. DOI: 10.3109/0142159X.2013.849798
10. Gierl MJ, Lai H. Evaluating the quality of medical multiple-choice items created with automated processes. *Med Educ.* 2013;47(7):726-733. DOI: 10.1111/medu.12202
11. Bennett RE. Technology for large-scale assessment. In: Peterson P, Baker E, McGaw B, editors. *International Encyclopaedia of Education*. 3rd ed. Oxford: Elsevier; 2010. p.48-55. DOI: 10.1016/B978-0-08-044894-7.00701-6
12. Bennett RE. The changing nature of educational assessment. *Rev Res Educ.* 2015;39:370-407. DOI: 10.3102/0091732X14554179
13. Seo DG. Overview and current management of computerized adaptive testing in licensing/certification examinations. *J Educ Eval Health Prof.* 2017;14:17. DOI: 10.3352/jeehp.2017.14.17
14. Krebs R. Prüfen mit Multiple Choice. Kompetent planen, entwickeln, durchführen und auswerten. Bern: Hofgrefe; 2019. DOI: 10.1024/85092-000
15. Bundesamt für Gesundheit. Eidgenössische Prüfung in Humanmedizin. Bern: Bundesamt für Gesundheit; 2019. Zugänglich unter/available from: <https://www.bag.admin.ch/bag/de/home/berufe-im-gesundheitswesen/medizin/berufe/eidgenoessische-pruefungen-universitaerer-medizinalberufe/eidgenoessische-pruefung-in-humanmedizin.html>
16. Meng-Hentschel J, Delmas C, Zurbuchen B, Wagner F, Schurter T. Erstellung von MC-Fragen für Prüfungen in der Medizin: individuell oder im integrierten Workshop. In: Gemeinsame Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA) und des Arbeitskreises zur Weiterentwicklung der Lehre in der Zahnmedizin (AKWLZ). Münster, 20.-23.09.2017. Düsseldorf: German Medical Science, GMS Publishing House; 2017. Doc353. DOI: 10.3205/17gma353
17. Krebs R. The Swiss way to score multiple true-false items: theoretical and empirical evidence. In: Scherpbier AJ, van der Vleuten CM, Rethans JJ, van der Steeg AF, editors. *Advances in Medical Education. Proceedings of the 7th Ottawa International Conference on Medical Education and Assessment*. Dordrecht: Kluwer Academic Publishers; 1996. p.158-161. DOI: 10.1007/978-94-011-4886-3_46
18. Angoff WH. Scales, norms and equivalent scores. In: Thorndike RI, editor. *Educational Measurement*. 2nd ed. Washington DC: American Council on Education; 1971. p.508-600.
19. Hofstee KW. The case for compromise in educational selection and grading. In: Anderson SB, Helmick JS, editors. *On Educational Testing*. San Francisco: Jossey-Bass; 1983. p.109-127.
20. Livingston SA, Zieky MJ. *Passing Scores, A Manual for setting standards of performance on educational and occupational test*. Princeton, NJ: Educational Testing Service; 1982.
21. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: The Danish Institute for Educational Research; 1960.
22. Norcini JJ, Boulet JR, Opalek A, Dauphinee WD. The relationship between licensing examination performance and the outcomes of care by international medical school graduates. *Acad Med.* 2014;89(8):1157-1162. DOI: 10.1097/ACM.0000000000000310
23. Tamblyn R, Abrahamowicz M, Dauphinee WD, Hanley JA, Norcini J, Girard N, Grand'Maison P, Brailovsky C. Association between licensure examination scores and practice in primary care. *JAMA.* 2002;288(23):3019-3026. DOI: 10.1001/jama.288.23.3019
24. Huwendiek S, Jung D, Schirlo C, Huber P, Balestra G, Guttormsen S, Berendonk C. The introduction of a standardised national licensing exam as a driver of change in medical education: A qualitative study from Switzerland. *Med Teach.* 2020;42(10):1163-1170. DOI: 10.1080/0142159X.2020.1798911
25. Boulet JR, Swanson DB, Cooper RA, Norcini JJ, McKinley DW. A comparison of the characteristics and examination performances of U.S. and non-U.S. citizen international medical graduates who sought Educational Commission for Foreign Medical Graduates certification: 1995-2004. *Acad Med.* 2006;81(10 Suppl):S116-119. DOI: 10.1097/00001888-200610001-00029
26. Huwendiek S, Mennin S, Dern P, Ben-David MF, Van Der Vleuten C, Tönshoff B, Nikendie C. Expertise, needs and challenges of medical educators: Results of an international web survey. *Med Teach.* 2010;32(11):912-918. DOI: 10.3109/0142159X.2010.497822
27. Huwendiek S, Hahn EG, Tönshoff B, Nikendei C. Challenges for medical educators: results of a survey among members of the German Association for Medical Education. *GMS Z Med Ausbild.* 2013;30(3):Doc38. DOI: 10.3205/zma000881
28. Huwendiek S, Reichert F, Duncker C, de Leng BA, van der Vleuten CP, Muijtens AM, Bosse HM, Haag M, Hoffmann GF, Tönshoff B, Dolmans D. Electronic assessment of clinical reasoning in clerkships: A mixed-methods comparison of long-menu key-feature problems with context-rich single best answer questions. *Med Teach.* 2017;39(5):476-485. DOI: 10.1080/0142159X.2017.1297525

Korrespondenzadresse:

Tina Schurter
Universität Bern, Institut für Medizinische Lehre, Abteilung für Assessment und Evaluation, Mittelstr. 43, CH-3012 Bern, Schweiz
tina.schurter@iml.unibe.ch

Bitte zitieren als

Schurter T, Escher M, Gachoud D, Bednarski P, Hug B, Kropf R, Meng-Hentschel J, König B, Beyeler C, Guttormsen S, Huwendiek S. *Essential steps in the development, implementation, evaluation and quality assurance of the written part of the Swiss federal licensing examination for human medicine*. *GMS J Med Educ.* 2022;39(4):Doc43. DOI: 10.3205/zma001564, URN: <urn:nbn:de:0183-zma0015649>

Artikel online frei zugänglich unter
<https://doi.org/10.3205/zma001564>

Eingereicht: 27.09.2021
Überarbeitet: 07.06.2022
Angenommen: 05.07.2022
Veröffentlicht: 15.09.2022

Copyright

©2022 Schurter et al. Dieser Artikel ist ein Open-Access-Artikel und steht unter den Lizenzbedingungen der Creative Commons Attribution 4.0 License (Namensnennung). Lizenz-Angaben siehe <http://creativecommons.org/licenses/by/4.0/>.