SMW Established in 1871
Swiss Medical Weekly
Formerly: Schweizerische Medizinische Wochenschrift
The European Journal of Medical Sciences

**Review article** | Published 3 December 2013, doi:10.4414/smw.2013.13897
**Cite this as:** Swiss Med Wkly. 2013;143:w13897

# The new licencing examination for human medicine: from concept to implementation

## Swiss licencing examination for human medicine

*Sissel Guttormsen*[a], *Christine Beyeler*[a], *Raphael Bonvin*[b], *Sabine Feller*[a], *Christian Schirlo*[c], *Kai Schnabel*[a], *Tina Schurter*[a], *Christoph Berendonk*[a]

[a] University of Bern Faculty of Medicine, Institute of Medical Education, Switzerland
[b] University of Lausanne Faculty of Biology and Medicine, Switzerland
[c] University of Zürich Faculty of Medicine, Switzerland

## Summary

A new Swiss federal licencing examination for human medicine (FLE) was developed and released in 2011. This paper describes the process from concept design to the first results obtained on implementation of the new examination. The development process was based on the Federal Act on University Medical Professions and involved all national stakeholders in this venture. During this process questions relating to the assessment aims, the assessment formats, the assessment dimensions, the examination content and necessary trade-offs were clarified. The aims were to create a feasible, fair, valid and psychometrically sound examination in accordance with international standards, thereby indicating the expected knowledge and skills level at the end of undergraduate medical education. Finally, a centrally managed and locally administered examination comprising a written multiple-choice element and a practical "clinical skills" test in the objective structured clinical examination (OSCE) format was developed. The first two administrations of the new FLE show that the examination concept could be implemented as intended. The anticipated psychometric indices were achieved and the results support the validity of the examination. Possible changes to the format or content in the future are discussed.

| **Abbreviations** | |
|---|---|
| ASM | anamnesis status management (history taking, physical examination, differential diagnosis, management plan) |
| CS | clinical skills examination |
| FLE | Swiss federal licencing examination |
| FOPH | Federal Office for Public Health |
| MCQ | multiple-choice question |
| MedBG | Medizinalberufegesetz / Federal Act on University Medical Professions |
| KK | Kommunikationskompetenzen / communication competencies |
| OSCE | objective structured clinical examination |
| IML | Institute of Medical Education; Medical Faculty, University of Bern |
| SAQ | short-answer question |
| SCT | script concordance test |
| SCLO | Swiss Catalogue of Learning Objectives for Undergraduate Medical Training |
| SP | standardised patient |
| USMLE | United States medical licensing examination |

***Key words:*** *licencing examination, examination formats, review, conceptual and scientific rationale, MCQ, OSCE*

## Introduction

Swiss medical education has a long tradition of promoting high quality and sustainable standards. This was also the aim with the release of the new Federal Act on University Medical Professions (MedBG) [1] and a new Swiss federal licencing examination (FLE). This paper presents the scientific rationale and the conceptual issues behind the development of this examination. The goals were ambitious: we aimed to develop a valid examination, specifically tailored to the Swiss medical education system, and encompassing the medical and educational practice of the country's five faculties of medicine. High standards for psychometric measures had to be balanced against feasibility and authenticity. In addition to its primary function as a quality control instrument for assessing knowledge and skills at the end of education, the FLE diploma also grants Swiss and foreign medical graduates (from countries outside the European Union or European Free Trade Area) the right to start postgraduate training or practise in Switzerland.

It was also clear that the examination would have implications for the design of the undergraduate medical curriculum. The central, key elements are presented below and explain the multifactorial character of the development process.

The MedBG addresses the quality of the medical professions. By way of example, Article 14 states that students shall possess broad knowledge, skills and adequate social competencies by the end of their education. Also, the Swiss Catalogue of Learning Objectives for Undergraduate Medical training (SCLO) [2] sets the context and specifies the objectives of the new FLE.

There are many stakeholders involved in the development and maintenance of a national licencing examination. The Federal Office for Public Health (FOPH) carries the responsibility for the legal and financial aspects of the FLE. The five Swiss faculties of medicine were actively involved both in defining the entire framework and in the content development processes. This was a purposeful decision, as the FLE should reflect the various educational systems and should also ultimately be accepted by the Faculties. The Institute of Medical Education (IML), University of Bern, provided expertise in the methodology of assessment and was committed to taking the lead in the central management of the FLE. The active process of developing the new FLE within this framework started in February of 2007, when the FOPH set up a national steering committee responsible for defining the strategic and political context of the FLE. A project group subordinate to the steering committee was established together with related examination specific subgroups that were in charge of the conceptual development and the operational planning and preparation of the specific examination format, content and logistics. In 2011, the steering committee was transformed into the national examination commission for human medicine and appointed by the Swiss Federal Council.

It was clear from the beginning that the FLE should have a high value with respect to validity, reliability, effects on the learning process, acceptance of assessment format, costs and practical feasibility [3]. A psychometrically sound and valid examination is "fair" for the students, because it minimises arbitrariness of the assessment outcome. The examination should be fair also in the sense that it reflects the content and learning objectives of the SCLO. It must also be defendable in that it must stand up to juridical appeals from candidates who fail. In order to achieve these aims the development and the decision-making process was supported by scientific discourse; pertinent research and literature was studied and the relevance to our aims and best practice – as well as applicability for the Swiss system – were continuously evaluated. The following questions were landmarks for the scientific discourse regarding the initial development:

1. *What are the assessment aims?* The FLE shall assess whether physicians possess the necessary knowledge, skills and social competencies as described in the MedBG at the end of their undergraduate education. How can these global aims be broken down into objectives that can be assessed?
2. *What assessment formats are appropriate for the FLE?* During undergraduate education, students are faced with many different assessment formats, written and oral examinations with theoretical and practical characteristics. Which aspects of the various formats have to be considered for the FLE?
3. *How can interdisciplinary knowledge and skills be assessed?* In line with the SCLO, the FLE follows an interdisciplinary paradigm. How can the notion of interdisciplinarity be optimally embedded and assessed?
4. *How should the FLE content development proceed in order to find common national foundation?* A challenge for the new FLE was to develop a coherent assessment content that reflected the training in all five faculties of medicine. How can the development process elicit a common understanding of the learning objectives listed in the SCLO?
5. *What concessions must be made when implementing a national licencing exam?* With our framework, the negotiation surrounding coherent and feasible assessment aims and formats was an important part of the development. Which ideas or assessment aims could not be realised? Why?

Thus, this paper is mainly devoted to the clarification of the above questions. The development of the FLE proceeded in three main phases, to which the structure of this paper can be mapped: (1.) concept development, (2.) content development and (3.) implementation. The landmark questions were a common thread throughout these development phases.

## Developing the examination concept

### Overall conceptual aims and conditions

The overall conceptual aims for the FLE were to create a feasible examination in accordance with internationally accepted assessment standards and thereby make a clear statement of the expected competence level of graduates at the end of their undergraduate medical education. Based on the initial reflections on the overall aims and legal aspects, the national steering committee set the following conditions for the further development of the future licencing exam: First, the FLE shall only cover deliberately defined aspects of knowledge and clinical skills. Second, these aspects should comprise applied clinical knowledge and practical clinical skills. Third, the FLE should be in the format of an objective, standardised examination in accordance with international standards. Fourth, the examination should be centrally developed but administered locally.

The decision to assess both applied clinical knowledge and practical clinical skills was based on the broad spectrum of competencies and objectives from the SCLO and the findings that the combined assessment of knowledge and skills better predicts candidates' readiness to enter professional life and advanced training than the assessment of only one of the two [4–6].

### Evaluating how to assess applied clinical knowledge

The assessment of knowledge by means of a written examination indicates that the candidates must possess a solid foundation of applied clinical knowledge, as well as practical clinical skills. Applied clinical knowledge can be tested with multiple-choice questions (MCQs) and short-answer questions (SAQs) in a licencing examination. In an MCQ examination the examinee is required to indicate true and false answers in a given set of alternatives or to choose an answer from the set [7, 8]. MCQ examinations are cost effective as they can be administered at a low cost per candidate. Alternatively, with SAQs the candidates must actively formulate the answers; hence it is commonly argued that this method is a more valid measure of existing knowledge [9]. However, this advantage is outweighed by the disadvantages of lower reliability per unit of testing time [10], of subjective factors from manual scoring of the answers [9] and a higher cost where there are large numbers of candidates as responses must be evaluated manually [10, 11].

The implementation of a MCQ examination would promise continuity both in regard to the previous federal examination, in which the MCQ format has been the standard for more than 30 years, but also with respect to undergraduate education as students at all Swiss faculties have gained much experience with this examination format. Many distinct and validated MCQ types have emerged over the years and overviews can be found in Case and Swanson [7] and Krebs [7, 12].

The script concordance test (SCT), a written standardised test based on a predefined set of selection alternatives, is an upcoming interesting alternative to classical written assessments. It probes a specific facet of clinical reasoning: the ability to interpret medical information under conditions of uncertainty [13]. Scoring reflects the degree of concordance of examinee judgments to those of a panel of reference experts [14]. During the initial development of the FLE this format was not considered because relevant experience among the faculties and IML was lacking; moreover, its implementation in a high-stakes examination has rarely been investigated.

### Evaluating how to assess practical clinical skills

The assessment of practical clinical skills for a high-stakes examination requires much effort in developing a standardised test in a realistic setting. Many less standardised assessment formats, such as unstructured (practical) oral examinations (e.g. the long case [15]) often applied in undergraduate medical education, were rejected mainly because such assessments cannot be adequately standardised. Moreover, unstructured oral examinations do not promise the assessment quality required for a licencing examination, as the validity and the reliability are low [16–18].

Standardised assessment of clinical skills is mostly carried out in the format of the “objective structured clinical skills examination” (OSCE). This examination format has become a gold standard in the context of high-stakes clinical skills examinations and is characterised by the use of standardised patients (SPs). This examination format comprises a circuit of stations in which candidates perform a series of different clinical tasks. Depending on the objective of each station, the tasks may include various clinical skills such as taking a focused history, performing a physical examination, providing counselling to a patient, deriving an accurate diagnostic hypothesis or proposing an appropriate management plan.

The OSCE format was first described by Harden et al. in the mid-1970s [19]. It has since been adopted around the world and has stood the test of time [20]. Various studies have demonstrated that OSCEs have good reliability coefficients [21, 22]. This has not only been shown for examinations at one single institution; comparable levels of reliability have also been found across multisite and multilanguage settings [23, 24]. Moreover, SP-based clinical examinations have demonstrated predictive validity for clinical performance, meaning that the scores achieved in an OSCE can predict the candidates’ professional performance. For example, Tamblyn and colleagues showed that scores on a standardised patient examination were significant predictors of competencies in consulting and prescribing in initial primary care practice [25]. Further, scores show a sustained relationship over 4 to 7 years with indices of preventive care and acute and chronic disease management [26]. Even more importantly, low scores achieved on national licencing examinations predict subsequent complaints to medical regulatory authorities [27]. The documented quality of the OSCE format has resulted in four large-scale certification and licencing examinations in Canada and the Unites States:

- objective structured clinical examination for family physicians in Quebec, Canada, since 1990 [28]
- objective structured clinical examination of the Medical Council of Canada qualifying examination part II, since 1992 [29]
- patient-based clinical skills assessment for foreign medical graduates as part of the United States medical licensing examination [30], since 1998 [23]
- clinical skill examination by the National Board of Medical Examiners, since 2004 [31].

An early pilot study was conducted in 2003 to evaluate the feasibility of a standardised patient-based practical examination as a possible component of the Swiss licencing examination. The outcomes from this pilot suggested that the scores of such an examination are reliable, valid and complementary to scores from written examinations. Moreover, preliminary experience was gained in how to set up a multi-institutional and interdisciplinary case development process [32].

The use of SPs instead of real patients is a topic of recurring debate. SPs for medical education were first described by Howard Barrows [33] and were initially used for neurological examination [34]. The use of SPs has been widely established since the early 1970s [35] and from 2002 became common in Switzerland [36]. The use of SPs for licencing examinations in the United States and Canada is well accepted and is now standard [30, 37]. Genuine patients have the advantage of showing more real pathologies compared than SPs [38]. However, real patients for a standardised licencing examination would negatively affect required levels of reliability and feasibility [39].

### Choosing the examination formats and specifying the assessment aims

The decision was taken to implement two complementary formats: an MCQ examination and a clinical skills (CS) examination in the format of a SP-based OSCE. It was decided that these formats would best be managed and developed centrally but locally executed in each of the five medical faculties. The problems listed in the SCLO represented pivotal starting points for both examination forms. The assessment aims were also specified accordingly:

- The new MCQ examination should assess applied clinical knowledge and the candidates’ ability to solve interdisciplinary problems. The notion of “applied clinical knowledge” also includes aspects of clinical reasoning. By comparison, the previous national MCQ licencing examinations as well as many of the undergraduate MCQ examinations used a discipline-related approach, assessing predominantly factual knowledge.

- The CS examination should assess whether the candidates can actually apply the clinical knowledge and skills necessary in order to enter residency. More precisely, the CS examination should focus on how candidates perform their clinical skills (e.g., take a history, conduct a physical examination) and how they communicate and interact with their patient. Hence, the CS examination focuses on observable clinical skills.

The decision to conduct MCQ and CS examinations had far-reaching consequences, as a number of competencies outlined in the SCLO cannot be tested in the FLE, for example, higher order competencies that go beyond knowledge and skills (such as acting professionally in a real-life setting). Because the FLE is carried out only after having successfully completed undergraduate education, the assessment of professional behaviour in daily practice and of longitudinal measures of performance was not considered and remains the responsibility of the medical faculties. It is intended that regular accreditation of study programmes will support student acquisition of such higher order skills.

## Developing the examination content

**Overall issues for examination content development**

To ensure FLE test content validity, a blueprint was set up in accordance with predefined criteria on which the MCQ and clinical cases were to be selected. The blueprint was published on the FOPH website together with the examination information and this was also provided to the candidates. It included two main dimensions and four secondary dimensions as listed below:

- Problems as starting points listed in the SCLO [2]. A problem is defined as a symptom, sign or test result of a patient with which the physician may be confronted. For example, a clinical case may be introduced to the candidate as follows: “You meet Mr X in Accident & Emergency, he complains of chest pain... Please take a focussed history and conduct a physical examination”. The main criteria for inclusion of symptoms are that they are either common or potentially dangerous and need to be acted upon quickly.
- Seven roles of the physician are part of the SCLO and adapted from the CanMEDS [40] model: medical expert, communicator, collaborator, manager, health advocate, scholar and professional. The main focus of the FLE lies with the role of the medical expert (MCQ, CS) and communicator (CS)
- Four secondary dimensions were included: (1.) setting – inpatient, outpatient; (2.) type of care – preventive, emergency, acute, chronic rehabilitation, palliative care; (3.) age – child, adult, elderly; (4.) gender – female, male, mixed

A format-specific development process for MCQs and CS was tailored to the quality and content requirements. Experts in medical education from the faculties took part in the respective development processes. Both examination formats and the Swiss FLE project were presented and discussed in depth at a workshop with experts from the National Board of Medical Examiners (USA) and the Medical Council of Canada in August 2010. This workshop and a series of pilot studies at the faculties allowed evaluation of the quality and process validity steps, which gave valuable insights for the finalisation of the examination development. Our experience from developing the examination is consolidated below.

**Multiple-choice question examination**

*Assessment dimensions*

The MCQ examination focuses on measuring applied clinical knowledge. Medical practice was reflected through the characteristics of the questions that required interdisciplinary problem-solving abilities. The evaluation of the various MCQ types convinced us that two forms of MCQ questions promised high levels of validity and reliability: the one-best-answer out of three to five answers item (type A) [41–44] and the multiple true-false item (type K prime) [43, 45]. In order to support the realistic application of clinical reasoning, rather than discipline-related knowledge testing, the MCQs were presented as a description of a concrete patient case (the so-called patient vignette [7, 46]. Each patient vignette included the various blueprint-dimensions, such as. the best diagnostic procedure (role of the physician: medical expert) of a 10 year-old boy (age, gender), brought to the family doctor (setting) with a knee injury (a starting point problem, type of care).

*Development of multiple-choice questions*

The development of the MCQ examination content followed cross-faculty engagement and interdisciplinary participation of professionals. Experienced clinicians – representing all specialties and all five faculties - were invited to MCQ writing workshops. In these workshops the clinicians were introduced to the design principles of good quality MCQ questions. This specifies that each question should follow the structure of a patient vignette, is relevant for medical practice and tests the application rather than simple retrieval of knowledge. Equipped with this information the clinicians then wrote the questions, individually or during the workshops.

*Consensual validation process*

A multipart revision process was then used to check the correctness of content, the relevance and the level of difficulty of each question in accordance with general practice and the SCLO. This process included revisions by inter-faculty and interdisciplinary groups of hospital-based clinicians as well as by representatives from general practitioners. Only questions passing successfully through the whole process were used in the FLE. Questions that did not reach consensus regarding content or did not meet the quality criteria were eliminated during this process; others were sent back to the authors with comments about how to improve them in accordance with the requirements. The latter would re-join the revision process once appropriately amended.

*Standardisation*

Standardisation of content and procedures was important throughout the whole development process, but particularly during the revision process. For formal quality, standardisation was achieved initially by a centralised formal review of the questions. This was done by assessment experts of IML in accordance with validated standards for MCQs [47]. Secondly, all candidates received standardised information about the examination (to be found on the FOPH website [48]. Thirdly, the candidates were offered an on-line accessible set of self-assessment questions that were similar in content, difficulty and time constraints to the questions used in the FLE [49]. Finally, each faculty received standardised instructions describing the requirements for the local organisation and execution of the examination, and the approved questions were translated into German and French. To ensure consistency the cross language translations were always conducted by the same person; both translators have a medical background and translated the questions into their native language. Medical professionals also reviewed the original and translated questions before they were included in the examination.

### Clinical skills examination

*Assessment dimensions*

The aim of the CS examination is to assess the practical clinical skills of the candidates, as practical observable skills during various encounters with SPs. Within a patient-candidate encounter the focus of the assessment lies within the following two main dimensions. The first dimension consists of content-specific aspects such as history taking, physical examination, diagnosis, management plan and counselling. The second dimension focuses on the communication aspects and the candidates' ability to engage in an empathic relationship.

*Development of the clinical skills stations*

The case development process started with the elaboration of adequate and valid topics. The starting point of the search for suitable topics was the chapter "problems as starting points for training" from the SCLO (comprising 277 problems). These topics had to fit the blueprint and needed to be feasible in an encounter with SPs. Once a specific topic was selected a content expert (senior clinician) at one of the five medical faculties was asked to write a case scenario. In a case development workshop this preliminary case scenario was further developed with the help of a second clinician (from both a different discipline and faculty) and a CS coach (a medical educationalist who had in-depth knowledge and experience with the CS test format). In these workshops not only was the case content *per se* developed, but also how the different aspects of the content should be weighed against each other. The operational description of the medical actions that candidates must undertake to solve a given problem was a major related part of the content development. The teamwork between the two experienced clinicians and the CS coach guaranteed that the quality aspects of the cases regarding both the content (does the case test relevant medical knowledge and skills?) as well as the test format (is the case appropriate for a 13-minute simulated patient-candidate encounter?) could be achieved. The practicality of cases was also tested during the case development workshops. This included role play with a SP as soon as the teams developed a case, which not only provided early input as to whether the case was feasible, but also identified necessary corrections.

*Consensual validation process*

The consensus-finding between the five medical faculties was a substantial element of the development process. Once the cases were drafted, they were reviewed by a national board consisting of members from all five medical faculties and general practitioners. The main tasks of the review board were (1.) to ensure that the developed test content was relevant with respect to the knowledge and skills accepted as necessary in order to act properly as a physician, (2.) to ensure that the degree of complexity matched the educational level of the candidates and (3.) to ensure that the test content was accessible and thus taught in a comparable way across all five medical faculties. The review board either accepted cases for inclusion in the CS examination, returned them for revision or they were declined. The validated cases were then handed on to the SP trainers.

*Standardisation*

SP trainers from the five medical faculties came together in meetings to familiarise themselves with and work through the validated cases. The SP trainers had to agree on exactly how the different roles should be portrayed. For challenging roles, videos were produced to clarify particular clinical scenarios, such as the extent of a neurological deficit to be portrayed. During these meetings discussions took place and finally consensus was reached about the equipment and its arrangement in the examination rooms, as well as questions concerning dress code and make up (i.e., skin rash) of SPs. Also for the examiner, standardised information and training meetings were held locally at all five medical faculties during which the role and duties of the examiner as well as how to rate the communication dimension were highlighted. The candidates received standardised written instructions for both examinations. In addition, the CS examination was exemplified with a video portraying the general procedure for an SP encounter. The translation of the cases was handled as described in the MCQ section above.

## Implementation

### Overall issues for the examination implementation

Pilot examinations in the CS format were conducted at most universities in 2010/2011. The main goal of these pilots was to establish feasibility (e.g., time per station) and to give the faculties and the students the chance to gain experience with the adapted format.

In advance of the examination, all candidates were informed about the content and purpose of the examination on the FOPH website. In addition, instructions for each examination were standardised and read to the candidates at the start of each examination day.

### Implementation of the multiple-choice question examination

The MCQ examination was administered locally at the five medical faculty sites simultaneously in two sessions. These sessions lasted 4½ hours each and were separated by one day. Each session contained 150 questions covering potentially all dimensions of the blueprint. The total number of 300 questions is regarded as necessary and sufficient to sample the examination content appropriately (as described in the SCLO and blueprint) and for a reliable measure of the candidates' knowledge (calculated on the basis of the former Swiss FLE and the Spearman Brown prophecy formula) and is in line with international standards (e.g. the licencing examination of the United States [30] and the qualifying examination of Canada [37]). As the USMLE allocates 90 seconds per question, the 108 seconds the Swiss MCQ allowed per question was considered adequate [42]. Each correctly answered type A question was rewarded with one point. For three correct answers in a type K question candidates were rewarded with 0.5 point, for four correct answers candidates received one point.

### Implementation of the clinical skills examination

The CS examination, like the MCQ examination, was administered locally at the five medical faculties. Depending on the number of candidates at the different sites, the examination was administered over two to four consecutive days. Each examination day had a different composition of clinical cases. The individual candidate was assessed in 12 stations, each with a 13-minute patient encounter session. Each examination day consisted of a different set of 12 stations, adding up to a total of 48 stations. All sites testing on a particular day used the same set of 12 stations. Each candidate was scheduled for a 3 hour 45 minute examination session (which included three 15-minute breaks and a 2-minute rotation time between the different stations). A total of 58 examination sessions with up to 14 students scheduled per session were administered. This set-up required that two to three sessions ran simultaneously and in parallel at each site during the four testing days.

Experienced clinicians who were recruited from the local faculty rated the performance of the candidates during the patient encounters. The aspects of history taking and physical examination, as well as the differential diagnosis and management plan (collated under the term ASM) were rated with a case-specific checklist. The communication skills (collated under the term KK) were rated instead with a uniform generic four-dimensional scale adapted from Hodges and Scheffer [50, 51]. A candidate’s score for a given station was composed of the sum of items checked by the examiner and the sum of the scores obtained on the four-dimensional KK scale, converted to a percentage. The total score from any station was a composite weighted score (total score = 0.75 * ASM + 0.25 * KK). All stations contributed equally to the total CS score. As each day had a different set of cases, the total scores were adjusted to a common mean (z-transformation).

All examiners involved in the CS examination had to participate in an orientation meeting and training session that highlighted the execution and scoring process of the CS examination.

## Overall outcome

### General information

The examinations in 2011 and 2012 were comparable with respect to format, content and results; hence detailed information is provided for the 2012 examination only: The number of candidates from the five Swiss medical faculties was 784 (MCQ) and 785 (CS). The number of candidates per faculty varied between approximately 120 and 240. Around 70% of the candidates took the examination in the three German-speaking faculties and 30% in the two French-speaking faculties. Additionally, there were a number of candidates with a foreign medical degree taking the examinations (43 MCQ and 16 CS) in order to obtain a Swiss medical diploma, as prescribed in the MedBG.

Results were calculated individually for all the candidates. The examination committee was mandated to set finally the pass/fail limit based on the calculations performed with the examination-related methods. Candidates who passed the examination received - in addition to their MC and CS scores – written information about how their performance related to their peers. Candidates who failed the examination received pertinent information about their performance in MC subscores as well as information about how their performance was rated in each CS case.

The results reported below serve only to illustrate proof of concept. This is not an empirical paper; therefore the details and reported data are kept to a minimum.

### Overall outcomes: multiple-choice question examination

The scores of the two examination sessions were totalled to generate the complete score per candidate. Only this total score was considered for the pass/fail decision. The reliability index (Cronbach alpha) for the examination items was 0.91 [52]. Table 1 shows the mean, standard deviation, minimum and maximum of the total examination score.

The national examination committee defined the pass score on the basis of two content-related methods [53, 54]. The pass rate of candidates from the five Swiss faculties was high (96.8–100%), whereas 67.4% of the foreign medical graduates passed the MCQ examination. The results were comparable between the faculties.

### Overall outcomes: clinical skills examination

Table 2 shows means, standard deviations, minimum and maximum of the total examination scores as well as the ASM and KK component scores collated over the four examination days. The Cronbach alpha values for the individual examination days were between 0.86 and 0.90. The total score is a composite weighted score. Overall the scores assessed a wide range of performance. However the scores achieved in the ASM component were consistently and substantially lower than the KK results.

In a subsequent step, the borderline regression method was applied to calculate the pass score [55]. The national examination committee then based their pass/fail decision on the calculated pass score. The pass rate of candidates of the five Swiss faculties was high (97.5–99.2%). In contrast, 50% of the foreign medical graduates passed the CS examination. It is important to note that the foreign medical graduates did not fail the CS examination because of insufficient communication (language) skills but because of low scores achieved in the ASM component, indicating a lack of applied clinical knowledge and skills.

**Intercorrelation between the clinical skills and multiple-choice question examinations and between clinical skills component scores**

Pearson correlation was used to determine the relationships between the MCQ scores and CS overall as well as the CS component scores. The moderate correlation (r = 0.52) between the CS and MCQ examination scores indicates that the two examinations measure separate and distinct competencies (that of course share a common ground) but are complementary in assessing the candidates' abilities. More interestingly, the low correlation between the KK component of the CS examination and the MCQ scores (r = 0.36) versus the moderate correlations between the ASM component of the CS examination and the MCQ scores (r = 0.51) support the construct validity of the two assessment formats.

## Discussion

The discussion addresses the five initial landmark questions in the context of the actual insights from the content and development process of the FLE.

The *assessment aims* for the FLE were to make a clear statement about the expected knowledge and skills level to be attained by the undergraduate at the end of their medical education to ensure both their readiness to start postgraduate training and the quality of medical professionals in Switzerland, in line with the newly introduced federal MedBG legislation. The development process resulted in the following operational aims in terms of measurable knowledge and skills: candidates should be able to handle clinical problems and take on the role of a medical expert and a communicator as presented in the SCLO; they should demonstrate applied clinical knowledge with sound clinical reasoning, as well as practical clinical skills, such as taking a history and performing a physical examination, and communication skills. The particular knowledge and skills are case-specific and good clinical practice is determined by clinical specialists in a thorough consensual process.

Among the many existing *assessment formats*, only a few are appropriate for a standardised national licencing examination. The formats under consideration had to be feasible and be acceptable to the five medical faculties. Furthermore, the selected assessment formats needed to have a record of convincing empirical evidence. The question of the assessment aims and the assessment format are interrelated: clarification of the assessment aims guides the selection of suitable assessment formats and conversely the selection of the assessment format should reflect the assessment aims. An important contextual issue for both examination formats was to represent credible medical problems in an appropriate environment in order to induce a minimal degree of immersion into the settings for the students (ecological validity). To achieve this, different considerations were applied: in the MCQ format, individual questions followed the description of a patient vignette to prompt realistic interdisciplinary clinical reasoning abilities; the CS examination adopted the same approach for the knowledge aspects. As the goal of the CS examination was to assess how candidates interact with a patient, much effort was invested in ensuring that the SP and the whole examination setting adequately represented the authentic environment.

The Swiss FLE comprises deliberately selected and weighted *assessment dimensions* and criteria that are regarded as necessary for postgraduate education in Switzerland. These dimensions are represented in the blueprint. The specifications in the blueprint composition represent the main criteria for the inclusion of an issue in the FLE. Thus the weighting of the different dimensions must be addressed regularly to ensure that the assessment dimensions continue to represent the intended purpose of the FLE. An important aspect of the MedBG is to set a clear structure for the FLE without imposing operational definitions. This therefore allows the development and refinement of the FLE over time, and guarantees that the FLE can be adapted to changing education and clinical practice frameworks.

Another important aspect of the assessment dimensions is the correlation between the different dimensions of the two examination formats. The correlations found between MCQ and CS results matches with the expectation of construct validity, as the correlations between the two formats are moderate on dimensions expected to be similar and low for aspects expected to be different: On one hand, the content-related aspects of, for example, history taking (CS) has more in common with the type of knowledge that is evaluated in an MCQ examination. On the other hand, and as expected, the ability to engage in an empathic patient relationship (CS) correlates weakly with the application of knowledge (MCQ). In other words, while the two examinations share some common ground, they were shown to be complementary with regard to various critical aspects. These results indicate that the two examinations do not overlap to such an extent as to make one of them superflu-

**Table 1**: Mean, standard deviation (SD), minimum and maximum as percentage of the total examination score.

| | Mean (%) | SD (%) | Minimum (%) | Maximum (%) |
|---|---|---|---|---|
| Score | 73.8 | 6.6 | 47.5 | 91.3 |

**Table 2:** Mean, standard deviation (SD), minimum and maximum as percentage of the total examination score – anamnesis status management (ASM) and communication competencies (KK) component scores.

| | Mean (%) | SD (%) | Minimum (%) | Maximum (%) |
|---|---|---|---|---|
| Total score | 73.0 | 6.1 | 44.8 | 87.9 |
| ASM score | 69.0 | 6.7 | 38.2 | 86.5 |
| KK score | 85.0 | 6.4 | 47.5 | 98.3 |

ous. These results are similar to the correlations found previously in the USMLE [56].

The management of the *content development process* was a key aspect underpinning generation of the FLE. It was clear that the different curricula from the five medical faculties had to be considered for the FLE. Part of the solution was to bring experts/clinicians from all the faculties together for the content development process. These meetings took place face-to-face and made fruitful and flexible discussions possible. We believe that the deliberate effort to build consensus was an important aspect of the successful FLE implementation. The starting point for many discussions was the Swiss Catalogue of Learning Objectives (SCLO). Although some limitations of the SCLO became evident, many lengthy debates regarding the desired relevance and depth of particular expert issues were avoided. Thus the SCLO guided content development and the consensus process enabled the experts to reach an accord in most cases. This was possible because, within the framework of the SCLO, the experts also had flexibility in how to formulate the assessment issues.

Simultaneously with the content development process, the educational deans met regularly with the steering group and the medical assessment experts. In this way the necessary acceptance for the product was in line with the development progress. Questions could be discussed regularly and the ownership by the faculties of the developing concept could be continually fostered. We also appreciated that close cooperation between the assessment experts, medical content experts and the decision makers (medical faculties and FOPH) was not only an aim, but also contributed to the success of the initiative.

In order to develop a standardised national licencing examination some *concessions* had to be made. The new FLE may give the impression of having sacrificed authenticity and flexibility for good statistical values, but according to the operational goals for the FLE, the examination meets expectations with regard to international standards. Nevertheless, some concessions were necessary for both examination formats. For the MCQs, it was apparent that the selection of one correct answer from a set of items does not reflect how patients present their problems in real life. However, given that the individual questions met the quality requirements, the proven feasibility and the reliability of the MCQ format outweigh the disadvantages in the context of a FLE. The CS examination is designed to compensate for the ecological validity of the MCQ format; however, other concessions are apparent: SPs lack real pathologies. It is clear that some patient groups cannot be portrayed by SPs, e.g., babies and small children, mainly for ethical reasons. As a consequence not all the necessary medical competencies can be assessed in the FLE. Thus, we also want to highlight that there are many higher order competencies that cannot be assessed with MCQ and CS examinations, or by means of a single point of measurement at all. Examples based on the CanMed roles include management competencies, professional attitudes and behaviour in daily work, collaboration within the medical team and with other professionals. Such skills must be assessed during undergraduate education. Therefore, the FLE does not by any means negate the responsibility of the universities to assess undergraduates in authentic situations during their education. Last but not least, in the context of a licencing examination it is not appropriate to provide situational feedback to the candidates about the qualitative aspects of their performance. Although this would be desirable from the point of view of the candidates, as well as for many examiners, it is important to keep in mind that the goal of the FLE is assessment *of* learning and not assessment *for* learning.

As an overall outcome of the FLE, the educational impact of this national licencing examination on students and the medical faculties should also be mentioned: In addition to guiding the students in what to learn, the FLE also directs the medical faculties as to what to teach and what opportunities to provide so that students can practice the essential clinical competencies. Through this process, we have experienced a growing nationwide awareness of, and fruitful discussion regarding, the necessary knowledge and skills physicians need to possess to successfully enter postgraduate training. It shall be an interesting topic of further research to investigate the influence of the FLE on the curricula of the five medical faculties.

## Outlook

After more than 4 years of development, the initial implementation of the new FLE examination has generated proof of concept. The intended psychometric indices have been achieved and the results support the construct validity of the examination formats. This is a good starting point and lays the groundwork for further development.

There is on-going discussion about how to evaluate and include additional clinical competencies in the FLE, and to address current concessions with more amenable solutions. Consequently the assessment concept can be extended with appropriate aims and assessment formats. Steps to improve aspects related both to the presentation of the clinical cases (external validity) and to the assessment formats are being dicussed. Aspects under consideration for future inclusion in the assessment include clinical reasoning strategies and more (real) pathological patterns with high-fidelity simulations in the FLE. High-fidelity simulators or computers carry the potential to present pathologies with sounds and video (e.g., heart murmurs, wheezing in a toddler, gait disorder in an elderly patient) more realistically. Multimedia-rich MCQs presented with computers were implemented on a pilot basis for the first examination in 2011. Computer-based assessment is one of the issues being addressed for the future development.

Following the scientific paradigm, future changes in the FLE format or content and the possible effects thereof will be followed up with data and continuously validated empirically. The question regarding the predictive validity of the FLE for later clinical work remains crucial. While predictive validity may not represent the only rational behind the FLE, such data would indicate how to guide curricular development, as well as subsequent improvements in the FLE. In this context, the educational impact of the FLE on the students' learning approach and on the curricula should also be investigated. Studies to obtain such information require carefully designed longitudinal research.

**Acknowledgements:**
The following national groups and experts contributed fundamentally to the success of the FLE:
Members of the national Exam Commission for the Federal Exam (2011–2013)
Members of the national 'Steering group Federal Licencing Exam' (2007–2011)
Members of the national 'Project group Federal Licencing Exam' (2008–2011)
Members of the national 'Working group MCQ' and the 'Working group CS' (2008–2013)
All the experts who contributed to the content development and implementation of the examinations

**Funding / potential competing interests:** No financial support and no other potential conflict of interest relevant to this article was reported.

***Correspondence:*** *Professor Sissel Guttormsen, Ph D, University of Bern Faculty of Medicine, Institute of Medical Education, Konsumstrasse 13, CH-3010 Bern, Switzerland, sissel.guttormsen[at]iml.unibe.ch*

## References

1 MedBG. Bundesgesetz über die universitäre Medizinalberufe; 2006 [cited 2013 Mar 16]. Available from: http://www.admin.ch/ch/d/as/2007/4031.pdf.

2 SCLO. Swiss Catalogue of Learning Objectives for Undergraduate Medical Training; 2008 [cited 2013 Mar 16]. Available from: http://sclo.smifk.ch.

3 Van Der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. Adv Health Sci Educ. 1996;1(1):41–67.

4 Wilkinson TJ, Frampton CM. Comprehensive undergraduate medical assessments improve prediction of clinical performance. Med Educ. 2004;38(10):1111–6.

5 Hamdy H, Prasad K, Anderson MB, Scherpbier A, Williams R, Zwierstra R, et al. BEME systematic review: Predictive values of measurements obtained in medical schools and future performance in medical practice. Med Teach. 2006:28(2):103–16.

6 Wenghofer E, Klass D, Abrahamowicz M, Dauphinee D, Jacques A, Smee S, et al, Doctor scores on national qualifying examinations predict quality of care in future practice. Med Educ. 2009;43:1166–73.

7 Case S, Swanson DB. Constructing Written Test Questions for the Basic and Clinical Sciences. Philadelphia: National Board of Medical Examiners; 1998.

8 Traub RE. On the Equivalence of the Traits Assessed by Multiple-Choice and Constructed-Response Tests. In: Bennett RE, Ward WC, editors. Construction Versus Choice In Assessment. Hillsdale (New Jersey): Lawrence Erlbaum Associates; 1993. p. 1–27.

9 Wesman AG. Writing the Test Item. In: Thorndike RL, editor. Educational Measurement. Washington: American Council on Education; 1971. p. 81–129.

10 Bennett RE. On the Meanings of Constructed Response, in Construction Versus Choice In Assessment. In: Bennett RE, Ward WC, editors. Hillsdale (NJ): Lawrence Erlbaum Associates; 1993. p. 1–27.

11 Schuwirth LWT, Van der Vleuten CPM. ABC of learning and teaching in medicine: Written assessment. Br Med J. 2003;326(7390):643–5.

12 Krebs R. Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung. 2004. Available from: http://www.iml.unibe.ch/dienstleistung/assessment_pruefungen/.

13 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):184–93.

14 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):189–95.

15 Norcini JJ. The death of the long case? Br Med J. 2002;324(7334):408–9.

16 Muzzin LJ, Hart L. Oral examinations. In: Neufeld VR, Norman GR, editors. Assessing Clinical Competence. New York: Springer Publishing Co; 1985. p. 71–93.

17 Swanson DB. A measurement framework for performance based tests. In: Hart IR, Harden RM, Editors. Further developments in assessing clinical competence. Montreal: Can-Heal; 1987. p. 13–45.

18 Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet. 2001;357(9260):945–9.

19 Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J. 1975;1(5955):447–51.

20 Williams RG. Have standardized patient examinations stood the test of time and experience? Teach Learn Med. 2004;16(2):215–22.

21 Boulet JR, en-David MF, Ziv A, Burdick WP, Curtis M, Peitzman S, et al. Using standardized patients to assess the interpersonal skills of physicians. Acad Med. 1998;73(10Suppl):S94–6.

22 Margolis MJ, Clauser BE, Swanson DB, Boulet JR. Analysis of the relationship between score components on a standardized patient clinical skills examination. Acad Med. 2003;78(10Suppl):S68–71.

23 Ziv A, Ben-David MF, Sutnick AI, Gary NE. Lessons learned from six years of international administrations of the ECFMG's SP-based clinical skills assessment. Acad Med. 1998;73(1):84–91.

24 Brailovsky CA, Grand'Maison P, Lescop J. A large-scale multicenter objective structured clinical examination for licensure. Acad Med. 1992;67(10 Suppl):S37–9.

25 Tamblyn R, Abrahamowicz M, Brailovsky C, Grand'Maison P, Lescop J, Norcini J, et al. Association between licensing examination scores and resource use and quality of care in primary care practice. JAMA. 1998;280(11):989–96.

26 Tamblyn R, Abrahamowicz M, Dauphinee WD, Hanley JA, Norcini J, Girard N, et al. Association between licensure examination scores and practice in primary care. JAMA. 2002;288(23):3019–26.

27 Tamblyn R, Abrahamowicz M, Dauphinee D, Wenghofer E, Jacques A, Klass D, et al. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. JAMA. 2007;298(9):993–1001.

28 Grand'Maison P, Lescop J, Rainsberry P, Brailovsky CA. Large-scale use of an objective, structured clinical examination for licensing family physicians. CMAJ. 1992;146(10):1735–40.

29 Reznick R, Smee S, Rothman A, Chalmers A, Swanson D, Dufresne L, et al. An objective structured clinical examination for the licentiate: report of the pilot project of the Medical Council of Canada. Acad Med. 1992;67(8):487–94.

30 USMLE, U.S.M.L.E.C.-. USMLE United States Medical Licensing Examination. [cited 2013 Mar 26]. Available from: http://www.usmle.org/.

31 De Champlain, A, Swygert K, Swanson DB, Boulet JR. Assessing the underlying structure of the United States Medical Licensing Examination Step 2 test of clinical skills using confirmatory factor analysis. Acad Med. 2006;81(10 Suppl):S17–20.

32 Vu N, Baroffio A, Huber P, Layat C, Gerbase M, Nendaz M. Assessing clinical competence: a pilot project to evaluate the feasibility of a standardized patient – based practical examination as a component of the Swiss certification process. Swiss Med Wkly. 2006;136(25–26):392–9.

33 Barrows HS, Abrahamson S. The Programmed Patient: A Technique for Appraising Student Performance in Clinical Neurology. Med Educ. 1964;39(8):802–5.

34 Barrows HS. Simulated patients in medical teaching. Can Med Assoc J. 1968;98(14):674–6.

35 Cleland JA, Abe K, Rethans JJ. The use of simulated patients in medical education: AMEE Guide No 42. Med Teach. 2009;31(6):477–86.

36 Schnabel KP. Simulation aus Fleisch und Blut: Schauspielpatienten. In: St. Pierre M, Breuer G, Editors. Simulation in der Medizin. Berlin, Heidelberg: Springer-Verlag; 2013. p. 115–9.

37 MCCQE. Medical Council of Canada Qualifying Examination (MCCQE). [cited 2013 Apr 26.03]; Available from: http://www.mcc.ca/en/.

38 Collins JP, Harden RM. The Use of Real Patients, Simulated Patients and Simulators in Clinical Examinations. AMEE Medical Education Guide No 13. Med Teach. 1998;20(6):508–21.

39 Hubbard JP, Levit EJ, Schumacher CF, Schnabel TG Jr. An objective evaluation of clinical competence. N Engl J Med. 1965;272(25):1321–8.

40 Frank J, editor. The CanMEDS 2005 physician competency framework. Better standards. Better physicians. Better care. Ottawa: The Royal College of Physicians and Surgeons of Canada. 2005.

41 Swanson DB, Holtzman KZ, Allbee K, Clauser BE. Psychometric characteristics and response times for content-parallel extended-matching and one-best-answer items in relation to number of options. Acad Med. 2006;81(10 Suppl):S52–5.

42 Swanson DB, Holtzman KZ, Allbee K. Measurement characteristics of content-parallel single-best-answer and extended-matching questions in relation to number and source of options. Acad Med. 2008;83(10 Suppl):S21–4.

43 Rogausch A, Hofer R, Krebs R. Rarely selected distractors in high stakes medical multiple-choice examinations and their recognition by item authors: a simulation and survey. BMC Med Educ. 2010;10:85.

44 Rodriguez MC. Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. J Educ Meas: Issues and Practice 2005;3–13.

45 Krebs R. The Swiss Way to Score Multiple True-False Items: Theoretical and Empirical Evidence. In: Schrepbier AJJA, et al., editors. Advances in Medical Education. Dordrecht: Kluwer Academic Publishers; 1997. p. 158–61.

46 Schuwirth LW, Verheggen MM, van der Vleuten CP, Boshuizen HP, Dinant GJ. Do short cases elicit different thinking processes than factual knowledge questions do? Med Educ. 2001;35(4):348–56.

47 "Institut für Aus-, W.u.F.n.I.f.M.L. Kompetent prüfen: Handbuch zur Planung, Durchführung und Auswertung von Facharztprüfungen. 1999. [cited 2013 Apr 26]; Available from: http://www.iml.unibe.ch/dienstleistung/assessment_pruefungen/.

48 BAG. Eidgenössisches Departement des Innern EDI. Bundesamt für Gesundheit BAG, Gesundheitsberufe 2013. [cited 2013 Mar 26]; Available from: http://www.bag.admin.ch/themen/berufe.

49 Self Assessment IML. 2013; Available from: http://self-assessment.iml.unibe.ch/.

50 Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Med Educ. 2003;37(11):p.012–6.

51 Scheffer S. Validierung des "Berliner Global Rating" (BGR) - ein Instrument zur Prüfung kommunikativer Kompetenzen Medizinstudierender im Rahmen klinisch-praktischer Prüfungen (OSCE). 2009, Dissertation, Medizinische Fakultät Charité, Berlin, Deutschland.

52 Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;(16):p.297–334.

53 Angoff WH. Scales, norms and equivalent scores, in Educational Measurement. Thorndike RL, Editor. Washington: American Council on Education; 1971. p. 508–600.

54 Hofstee KWB. The case for compromise in educational selection and grading, In: Anderson SB, Helmick JS, editors. On Educational Testing. San Francisco: Jossey-Bass; 1983. p. 109–27.

55 Kramer A, Muijtjens A, Jansen K, Düsman H, Tan L, van der Vleuten C. Comparison of a rational and an empirical standard setting procedure for an OSCE. Objective structured clinical examinations. Med Educ. 2003;37(2):132–9.

56 Harik P, CB, Grabovsky I, Margolis MJ, Dillon GF, Boulet JR. Relationships among subcomponents of the USMLE Step 2 Clinical Skills Examination, the Step 1, and the Step 2 Clinical Knowledge Examinations. Acad Med. 2006;81(10):S1–4.