Language Matters: Development of an Objective Structured Language Test for Foreign Physicians – Results of a Pilot Study in Germany

Abstract

Objective: To develop a scientifically sound and standardized medical language examination for the State of Bavaria according to the requirements set forth by the 87th Conference of State Health Ministers. This *Sprachtest für Ausländische Mediziner* (SAM, Language Test for Foreign Physicians) ought to become part of the licensing procedure for foreign physicians in Germany. Using testing stations that are situation-based, it will assess medical language competence and communication skills at the proficiency level of C1.

Methods: Case scenarios for four mini-interviews of 10 minutes each were developed. For the written part of the exam, consisting of two separate testing stations with a combined duration of 40 minutes, one video of a physician taking a patient's history and one annotated set of laboratory results were developed. Based on the analysis of existing scientific literature as well as real-life examples, features and characteristics of professional medical language were identified. This served as the basis for the development of itemized rating scales for each of the testing stations. The exam was validated in three simulated trial runs. Each run was video-recorded and subsequently graded by a team of test-raters.

Results: 19 participants took part in the three trial runs. A benchmark (gold standard) could be set for 18 of these. A ROC-analysis yielded an AUC-value of .83. This confirmed the predictive quality of the SAM-test. The reliability of the SAM-test could be calculated for only ten participants. The internal consistency, calculated with the use of Cronbach's Alpha, was .85. The pass/fail mark was calculated based on the Youden-Index and yielded a result of >60%.

Conclusion: The SAM-test presents a statistically valid medical language examination with a high level of objectivity. As required, it tests language proficiency at the level of C1 and uses authentic communication scenarios within a standardized test setting. Additional studies with larger test samples will help to further validate this test and thus guarantee a higher degree of reliability.

Keywords: medical language, exam, foreign physicians

Holger Lenz[1]
Ansgar Opitz[2]
Dana Huber[3]
Fabian Jacobs[1]
Wolfgang Gang Paik[4]
Jörg Roche[5]
Martin R. Fischer[1]

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

2 LMU München, Lehrstuhl für Empirische Pädagogik und Pädagogische Psychologie, München, Germany

3 LMU München, (ehem.) Institut für Deutsch als Fremdsprache, München, Germany

4 LMU München, Medizinstudierender, München, Germany

5 LMU München, Institut für Deutsch als Fremdsprache, München, Germany

1. Introduction

"Anyone who focuses only on the slightly increasing number of physicians closes his eyes to the whole truth. In reality, the gap between demands for medical care and the capacities to give it continues to widen steadily." [1]. Thus commented the president of the German Medical Association, Frank Ulrich Montgomery, on the nationwide statistic for physicians from 2016. For some time now, buzzwords such as "shortage of physicians" and "shortage of skilled workers" have been circulating through the public discourse on health policy [2]. More and more physicians from foreign countries continue to close the gap: within the past five years, their number in Germany has nearly doubled. It reached a record high in 2016 with a total of 41.658 [3].

As they go through the process of integrating into their everyday professional lives, however, foreign physicians are confronted with a number of technical, administrative and cultural challenges that are partly driven by a lack of language proficiency. Insufficient or deficient communication skills often result in a lower quality of treatment, lower levels of patient satisfaction as well as intercollegiate conflicts and therefore pose a significant threat to patient safety. In extreme cases, failure to communicate successfully can be the decisive factor whether a patient dies or lives [4], [5], [6], [7], [8]. Proficient communication

that clears misunderstandings and prevents them is a vital element of medical practice [9].

Therefore, the 87th *Gesundheitsministerkonferenz* (GMK, Conference of State Health Ministers) resolved to make obligatory a nationwide language exam for healthcare professionals and, at the same time, specified a number of minimum requirements. Those requirements include one simulated conversation between the healthcare professional and a patient; the composition of a document in written form as it commonly occurs in the daily routine of the healthcare professional; and a conversation with a member of the same profession. Each part should last 20 minutes [10] (cp. Abbildung 1). To this date, there are no common standards for the theoretical and methodological framework of this test. The formal requirements as established by the GMK refer mostly to the language level C1 when used as an expression of language in a professional context or setting.

It is true that the requirements did create the necessary framework for higher standards of language proficiency. The responsibility to guarantee language exams of high quality, however, was shifted to the individual states. As can be seen from an overview issued by the *Marburger Bund*, the conceptualization of the actual exam varies greatly from state to state [11]. The lack of a common national exam, however, causes the risk of so called "exam tourism". This means that foreign medical professionals will try to pass the examination in those states, in which the exam is supposedly easier to pass than in other states. In 2016, the state of Bavaria represented by the *Staatsministerium für Gesundheit und Pflege* (StMGP, State Healthcare Department), commissioned an interdisciplinary research team from the medical faculty, the Institute for Medical Education, the Department of German as a Foreign Language and from psychometrics at the *Ludwig-Maximilians-University* (LMU) with the development of a valid, reliable, fair, authentic, objective and viable *Sprachprüfung für ausländische Mediziner* (SAM, Language Test for Foreign Physicians).

In the Anglo-American world, the Australian test model can be considered as the leading model, since it, too, is based on similar scientific and methodological standards [12]. An analysis of this model, however, has shown that international test models can be used only as a general guideline. Even the Australian model does not meet all scientific criteria of test development [13], [14]. This made it absolutely necessary to create an independent methodological foundation for the SAM-test. This article outlines the design of the SAM-test and discusses the initial testing phase along with its results.

2. Project Description and Methodology

Considering the requirements as stated in the resolution of the 87th GMK, the research team developed a design concept that primarily focused on meeting the quality criteria of objectivity, reliability, validity and authenticity [10], [15], [16], [17].

2.1. Design of the examination

The schematic design of the exam can be seen as displayed in figure 1. *Taking A Patient's History and Patient Consultation Before A Surgical Procedure* were chosen as topics for the part of doctor-patient-communication. When *Taking A Patient's History*, the examination candidate has to verbally obtain information relevant to the patient's medical condition, allow adequate time for the patient to report about his symptoms and create an atmosphere of respect. At the same time, the candidate's ability to understand spoken language is tested. During the *Patient Consultation*, the focus shifts towards the transmission of information. The physician needs to explain the process of the upcoming surgical procedure, point out potential risks and give detail instructions about post-surgical precautions and measures. Focus of this test section is the use of vernacular (avoiding technical medical terms), ascertaining that the patient has understood all relevant information as well as verbally and nonverbally expressing empathy toward the patient's questions and concerns.

A prototypical communication situation for the part 'intercollegiate communication' is the presentation of a patient's case to others: *Relating A Patient's History* and condition to the senior physician. In this section of the test, which is also based on a simulated case scenario, the examinee has to use technical medical language and terminology to demonstrate successful communication with a colleague (here: a senior physician). Both the act of relaying information as well as stating clear and clarifying questions should be done in a concise way and with accuracy.

In contrast with other medical language exams in Germany and as required by the GMK, the SAM-test also includes the examination of communicative proficiency when it comes to conversations between physicians and professionals from other healthcare professions [10]. The *Instructive Conversation With A Nurse* was thus chosen as a typical scenario for this type of communication format. In this test section, clearly stated instructions have to be given to a nurse. This should also happen while using appropriate technical language and terminology in a respectful atmosphere.

For the written part of the exam, an analysis of 200 physician letters from the fields of surgery and internal medicine at the university hospital of the LMU revealed that physician letters generally consist of four structural elements. Two of those, *Case History and Reason for Admission* and *History and Treatment Plan* were included in the SAM-test because of their high level of difficulty. The written part tests the examinee's ability to receive and process language input along with the ability of making verbal expression in written form.

Cases from the subject areas of general medicine, internal medicine and surgery were chosen for the case scenarios.

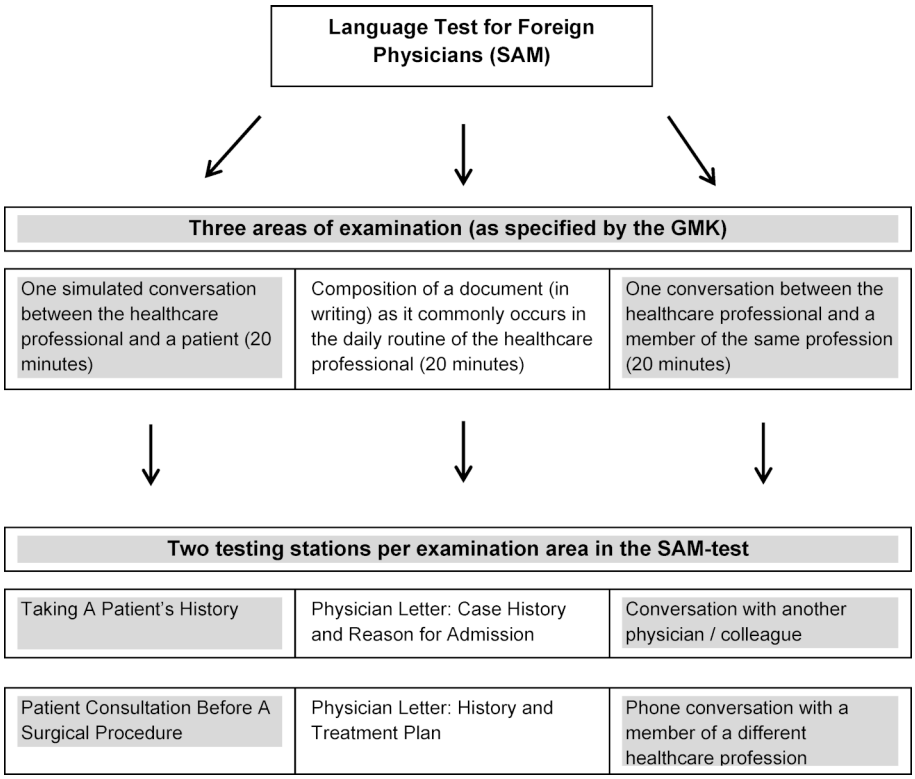

Figure 1: Schematic design of the language test for foreign physicians

These areas generally correspond with the content areas of a subsequent examination that foreign physicians from countries which are not members of the European Union have to take in order to demonstrate their medical-technical know-how at the level of the 3rd State Examination before receiving their license to practice. Thus, setting the focus on these subject areas can be seen as justified regardless of the personal specialty area of each candidate.

Case scenarios were kept as general as possible in order to avoid focusing the exam too much on content specific to one area of medical practice. During the *Patient Consultation*, for example, examinees deal with scenarios from common surgical procedures such as a thyroidectomy or a tonsillectomy.

2.2. Format of the examination

The OSCE-format (Objective Structured Clinical Examination) was chosen for the SAM-test to meet the real (authentic) demands of everyday professional practice and at the same time create conditions comparable to those existing for medical students, who have to prove their medical know-how at a university. According to Miller, OSCE-exams offer the opportunity to not simply reproduce knowledge, but to show what one has learned in a practical, context-driven setting [17]. From the viewpoint of medical educators, OSCEs have established themselves as reliable and valid instruments when it comes to testing clinical-practical knowledge [18]. Brandes and Bagnasce et al. have further shown that OSCEs are well suited as a methodological setting for measuring communication skill levels in cultural and professional contexts [19], [20].

Analogous to the OSCE-concept of multiple short test scenarios with a length of five to ten minutes, the SAM has been designed with two testing stations of ten minutes each for each one of the two areas that examine oral proficiency (cp. figure 1). This leads to an increase in reliability, since the performance of the examinee can be observed four times in four different contexts. Additionally, ten minute scenarios more realistically represent the time frame available to physicians during their daily routine, which therefore increases the level of authenticity of the test.

2.3. The problem of interdependent testing station results

Current medical language examinations often use one case scenario throughout the entire test. From a psychometric point of view, however, this concept presents challenges: if *one* case scenario is used throughout the entire exam, this creates a dependency between the assessment items for each testing station of the exam: the results in one area no longer depend solely on the performance in that area, but also on the performance in preceding test areas [15].

Additionally, the "one case scenario" model leads to a drastic reduction of *fairness*: if the candidate is accidentally tested in an area that s/he is especially familiar with due to former experience or past medical education, his or her test performance is automatically better. Finally, using a model in which test areas are independent of each other alleviates the exchange of case scenarios that need to be removed from the exam due to repeated use:

if a test consist of multiple case scenarios, it is possible to compare the level of difficulty of a new case scenario with the level of difficulty of existing ones; if, however, a test consists of only one case scenario, the exchange of that one scenario automatically leads to the exchange of the entire test. This, however, makes it impossible to assess the level of difficulty of the new case scenario in relation to other case scenarios. Therefore, different case scenarios with multiple testing stations have been used for the SAM-test.

2.4. Implementation and assessment

Every language examination that aims at testing the examinee's productive and receptive language abilities has to create communication situations that are as realistic as possible (authentic) and as reproducible as possible (objective and fair). This ensures that all candidates are tested within the same communicative contexts. To create such standardized communication settings, the SAM-test makes use of trained actors for the roles of the "patient" and the "nurse". The role of the senior physician is filled by an actual, real-life physician.

Both the actor simulating the patient and the real-life physician attend multiple training units to prepare for their roles. The main emphasis of the training units is to create standardized test settings (objectivity, fairness) and to evoke language patterns that are specific to each case scenario. A script for the simulated patient with detailed instructions and additional questions was developed.

Current medical language examinations in other German states assess the examinee's performance in a synchronous way: a group of raters present in the testing room observe the candidate's performance and evaluate it, often with the help of standardized assessment sheets. Synchronous assessments of oral performance, however, are problematic in many ways: what is expressed verbally is fleeting by nature and cannot be reviewed; assessment is also made "out of the (ongoing) situation" and raters are often participants in the communication situation.

Asynchronous assessment with raters who are not part of the communication situation and who only assess the oral parts of the exam, however, allows for repeated, independent and standardized listening to the candidate's performance and thus increases the objectivity of results. Therefore, oral test parts are video-recorded in the SAM-test. This method of testing and performance evaluation, called VOSCE (Video-Recorded Objective Structured Clinical Examination), has successfully proven to be a feasible, reliable and valid method to assess communication ability in other medical contexts [21], [22], [23]. Since storing and accessing recorded test data can be problematic in view of strict data protection and privacy laws, a special software program was developed. This program records each performance through an external camera attached to a laptop computer and stores the recorded, pseudonymized data on a password-protected server, which allows for secure access of recordings by the team of test raters at a later point in time. This team consists of one physician and one linguist with a background in German as a Second Language theory and test methodology. An itemized rating scale was developed for each testing station (*History Taking, Patient Consultation*, etc.). For each item, the rater must choose between three different possibilities: "Standard was met", "Standard was not met" and "Not sure". The option "Standard was met" is equivalent to one point, the option "Standard was not met" to 0 points and the option "Not sure" to 0.5 points. All items are categorized according to the typical structure of the professional language in use, the linguistic pattern or style, the behavior in the communication situation as well as the global impression of the performance in the communication situation as a whole. Each rating scale consists of between 11 and 17 items, which adds up to a total of 83 items for the SAM-test (cp. table 1). An example of a rating scale for the subpart *History Taking* can be found in attachment 1.

A supplementary sheet for each rating scale explains the intended use of the items and gives case-specific examples. This complies with the requirements of the Association of Language Testers in Europe (ALTE) for language test assessment procedures [24], and increases the probability of a standardized and consistent rating process. Additionally, the test-developers provided a training session (ca. one hour) for each new team of test-raters in order to explain the rating process and answer any pending questions.

Rating of test performances first occurs individually. Afterwards, the team of raters has to agree unanimously whether a candidate passes or fails the test. After assessing the test performance individually, raters compare their results and must reach a consensus for any diverging assessment of the rating scale items. The cumulative result of all six testing stations finally decides whether a candidate passes the test or not.

3. Pilot Testing of the SAM-Test

3.1. Implementation

During the pilot testing phase, the SAM-test was validated in three simulated trial runs. A total of 19 candidates participated in the trial runs. These came either from the pool of international medical students at the LMU (n=10) or from the pool of international physicians who live in Germany, but do not yet have their license to practice medicine (n=9). With the help of these simulations, it could be determined how feasible it was to implement the design of the SAM-test. Additionally, the results were used to determine to which degree rater evaluations of performances are in agreement, to measure reliability, to evaluate the prognostic ability of the test and to determine the pass/fail mark. In order to determine the pass/fail mark and to understand the prognostic ability of the test, a benchmark (gold standard) was used: In addition to the (regular) assessment of participants'

Table 1: Reliability values of the individual testing stations of the SAM-test

Testing Stations (In parenthesis: number of items, number of participants used to calculate the value)	Reliability (Cronbach's alpha)
Physcian Letter - Case History and Reason for Admission (17 Items; N = 18)	.903
Physcian Letter - History and Treatment Plan (13 Items; N = 18)	.929
Conversation With A Patient - Taking A Patient's History (16 Items; N = 14)	.594
Conversation With A Patient - Patient Consultation Before A Surgical Procedure (12 Items; N = 14)	.224
Instructive Conversation With A Nurse (12 Items; N = 13)	.675
Relating A Patient's History (11 Items; N = 15)	.911

performance in the SAM-test by a team of test raters, an expert team consisting of two professionals from the subject areas "Medicine" and "German as a Foreign Language" with many years of experience in assessing communication performances joined the rating process. These experts used a global rating system to determine whether candidates had reached the minimal requirement of the C1 language level. Comparing the itemized results of the regular rating team with the assessment of the two expert raters (which was used as the gold standard) allowed for evaluating the quality of the SAM-test as well as for setting the pass/fail mark.

3.2. Results

It is best to use Cohens Kappa to determine as to which degree the two raters' performance evaluations are in agreement. This indicates to what extent the consensus of the two raters is higher when compared with a set of randomly generated evaluations. Possible values range from 0 to 1. Through the use of training sessions, the SAM-team was able to raise the consensual value of evaluations from .49 to .72. At the end of the pilot testing phase, the percentage of consensual evaluations was at 88% (cp. to 80% at the outset).

Because of missing data as well as minor adjustments of the rating scales between the first and subsequent trial runs, the reliability of the overall SAM-scale could only be calculated for ten candidates and 81 items. The internal consistency of this set of 81 items, calculated with the use of Cronbach's alpha, was .85. The reliability values for each testing station (for which there is more data) can be seen in table 1.

On average, all candidates fulfilled M=55% (SD=22%) of the 83 items of the six rating scales. A benchmark (gold standard) could be set for 18 candidates. Five were rated as reaching the minimum qualification of the C1 language level. The performance result of M=69% (SD=19%) of these five candidates was higher than the performance result of those who did not achieve the minimum requirements according to the benchmark (gold standard) (M=46%, SD=14%). To accurately examine the prognostic ability of the SAM-test (in relation to the eligibility of the candidates), a ROC-analysis was used (receiver operating characteristics) [25]. This analysis determines to what extent the performance in a test corresponds with the "actual" proficiency of the candidate (represented by the benchmark). The global quality level of the test can thus be quantified by using the AUC-value (area under the curve). The AUC-value can range from 0 to 1. An AUC-value of 0.5 means that the test is no better than mere chance in determining which candidate is qualified and which one is not. An AUC-value of 1 means that the assessment of all candidates is correct. For the SAM-test, an AUC-value of .83 was determined. According to current test methods, this shows a strong effect and emphasizes the prognostic quality of the SAM-test [26].

Moreover, the pass/fail mark was determined with the help of the ROC-analysis. To do this, the Youden-Index was used [27]. This index combines the sensitivity (the number of candidates who are qualified and are correctly identified as such by the test) and specificity (the number of candidates who are not qualified and are correctly identified as such by the test) of the test into one single value. Higher values are desired. A pass/fail mark of 50% produced a value of .49. At this mark, the values of sensitivity and specificity were at .80 and .69 respectively. The PPV (positive predictive value; the probability that a candidate is truly qualified once the pass/fail mark has been reached) lies at .50 for this threshold, and the NPV (negative predictive value; the probability that a candidate is truly unqualified if the pass/fail mark is not reached) at .90. A pass/fail mark of >60% results in a Youden-Index value of .52. Even though the sensitivity drops to .60, the specificity value rises to .92. The PPV is .75 and the NPV .86. If the Youden-Index is used as a criterion and one assumes that the highest priority of any medical language examination is to prevent possible damage to the general public, a more conservative threshold of >60% should be used. In this context, "conservative" means that a

candidate whose performance falls in the borderline area between qualified and unqualified is deemed as unqualified. The data of the trial runs even allow for the possibility to raise the pass/fail mark to 70%. Without loss of sensitivity, this would result in a rise of the specificity value to 1. However, since the distribution of data suggests that at such a pass/fail rate, the sensitivity value would drop off once larger data sets are used, and since the specificity value at the >60% mark is already very high (.92%), a pass/fail mark of >60% is suggested for the SAM-test. Table 2 represents an overview of the most important statistical results.

Table 2: Overview of the most important statistical data of the SAM-test

Statistic	Value
Total Reliability Value of the Test[a]	.85
Reliability Values of the Testing Stations[a]	.22 - .93
Interrater Objectivity[b]	.72
AUC	.83
Sensitivity[c]	.60
Specificity[c]	.92
PPV[c]	.75
NPV[c]	.86

Notes:
a) measured as Cronbach‘s alpha;
b) measured as Cohen‘s Kappa;
c) Value when considering the recommended pass/fail mark of >60%

4. Discussion and Conclusion

Good results were achieved especially in the areas of fairness, authenticity and objectivity. In this context, it is important to again emphasize the importance of coaching all actors who participate as simulated patients in the communication situation. Only if the simulated patient acts in a consistent matter towards each and every candidate can a reproducible test environment be guaranteed. The resulting increase in test-objectivity in turn has a positive effect on the reliability and validity of the test. Inversely, the low reliability value of the testing station *Patient Consultation* could possibly be explained by referring to the occasional but unintended observation of simulated patients giving assistance during the communication situation. It is possible that simulated patients (who do not have a background in medicine) give cues to weaker candidates out of a feeling of empathy. This would reduce the systematic variance of results and thus affect reliability. This and other data collected within the context of this project about the respective peculiarities and challenges that both simulated patients and exam candidates encounter within each communication situation of the test can therefore serve as an initial basis for the development of a standardized, scientifically verified training method.

Another strength of the SAM-test lies within the concept of evaluating test performances in an asynchronous manner. Test raters who experience the communication situation “live” or are even part of the communication situation themselves increase the risk of introducing *bias* into the rating of the candidate's performance. The model of asynchronous assessment of test performance used in the SAM-test contributes to a fair and objective evaluation of all examinees und thus reduces the risk of legal complaints on the part of exam candidates.

The validity values of the SAM-test based on the ROC-analysis of data from the pilot testing phase are promising. This is especially so considering that, according to the benchmark (gold standard), the rate of qualified candidates was low, which in turn complicates the process of identifying qualified candidates. When analyzing the results, it is furthermore important to bear in mind that half of the participants in the trial runs were foreign students. Since students have less experience and knowledge than experienced physicians, it is possible that this contributed to a distortion of the collective performance results of all candidates. Within the sample group of experienced physicians, the rate of qualified candidates should thus be higher. It is further necessary to take into consideration that the relatively small sample group from all three trial runs implies a high level of uncertainty of all test parameters. A more systematic validation of the test is therefore absolutely necessary. For example, the overall good validity of rating scales during the pilot testing phase and subsequent performance assessment is at odds with the unsatisfactory reliability values of the rating scales for two testing stations (*Patient Consultation* and *Instructing a Nurse*). Future trials that intend to reduce the deficiency of above mentioned scales and aim at increasing the psychometric quality of all scales could therefore especially benefit from trial samples of larger size and consisting of a more homogenous group of candidates respective their language ability and proficiency. A more precise measurement of the reliability value of the SAM-test would thus be a natural consequence of a larger sample size.

Another weakness of the SAM-test lies in the initial investment costs needed for setting up the test environment (software program and training of the simulated patients and raters). The longer the SAM-test runs, however, the more should its strengths serve to offset this disadvantage.

Further action is needed regarding the distribution of the number of items for the rating scales. The item number for the scales of each individual testing station varies between 11 and 17. In order to give equal weight to the scale of each testing station, a retroactive adjustment is recommended to avoid the need to artificially increase or decrease the number of items. Before calculating the total sum value for the entire test, the point value achieved in each of the six testing stations would have to be multiplied by different coefficients so that the candidate can achieve exactly $^1/_6$ of the total maximum points in each testing station.

5. Outlook

To this date, the SAM-test represents the first and only scientific concept of a medical language test within Germany. In addition to the parameters set out by the GMK, quality standards of test and measurement theory such as objectivity, reliability, validity, authenticity, fairness and feasibility were closely adhered to as guiding principles of design and implementation. The SAM-test is also currently the only medical language examination in Germany that includes the aspect of inter-professional communication. In addition to the introduction of the communication setting between a physician and a nurse, it is conceivable to include further situations that produce inter-professional communication situations. In view of the goal to create and maintain a scientific and robust examination, it must be noted that further simulated trials are necessary.

It is further recommended to compare the SAM-test with other examinations to see how they measure up to the quality standards of test and measurement theory. Only then can the goal of a unified national exam, which reliably tests foreign physicians at the language level of C1 and thus guarantees patient safety, finally be reached. At the time of this writing, one additional comparative study with the goal of validating examinations currently used in the state of Bavaria is being planned. It is the professed aim of the test developers to see the SAM-test being used in the foreseeable future and thus to contribute to the lasting improvement of current methods of testing – not only in the State of Bavaria.

Acknowledgements

For the sustainable support of the project, we would also like to thank Prof. Dr. med. Matthias Siebeck, Department of General, Visceral, Transplantation, Vascular and Thoracic Surgery of LMU Munich

Funding

We would like to thank the Bavarian State Ministry for Health and Care (StMGP) for the support of the project under grant number G32g-G8517.1-2015/5-91.

Competing interests

The authors declare that they have no competing interests.

Attachments

Available from
http://www.egms.de/en/journals/zma/2019-36/zma001210.shtml

1. Attachment_1.pdf (141 KB)
   Language Test for Foreign Physicians

References

1. Bundesärztekammer. Die Schere zwischen Behandlungsbedarf und Behandlungskapazitäten öffnet sich. Berlin: Bundesärztekammer; 2017. Zugänglich unter/available from: http://www.bundesaerztekammer.de/presse/pressemitteilungen/news-detail/die-schere-zwischen-behandlungsbedarf-und-behandlungskapazitaeten-oeffnet-sich/
2. Bundesärztekammer, Kassenärztliche Bundesvereinigung. Dem deutschen Gesundheitswesen gehen die Ärzte aus! Studie zur Altersstruktur- und Arztzahlentwicklung. 5th ed. Berlin: Bundesärztekammer und Kassenärztliche Bundesvereinigung; 2010. Zugänglich unter/available from: http://www.kbv.de/media/sp/Arztzahlstudie_2010.pdf
3. Bundesärztekammer. Ärztestatistik zum 31. Dezember 2016. Berlin: Bundesärztekammer; 2017. Zugänglich unter/available from: http://www.bundesaerztekammer.de/fileadmin/user_upload/downloads/pdf-Ordner/Statistik2016/Stat16AbbTab.pdf
4. Wichmann R. Weitere Zehn Jahre Warten hilft nicht. Praxisguide D Krankenhaus. 2015;(1):14-15.
5. Karimi P, Rudenko O. Am Anfang verstand ich null. Praxisguide D Krankenhaus. 2015;(1):20-21.
6. Arndt J. Sprachbarrieren im Krankenhaus – Wenn dem Arzt die Worte fehlen. Pneumologie. 2016;70(9):564-566. DOI: 10.1055/s-0042-114156
7. AG Leipzig. Aufklärung durch einen Arzt, der die deutsche Sprache nicht beherrscht. MedR. 2003;10:582-583.
8. Roche J. Zur Frage der Deutschkenntnisse. Sprache Beruf. 2014;7:316-318.
9. Schröder H. Theoretische Aspekte der Arzt-Patienten-Interaktion. In: Witt C, ed. Der gute Arzt aus interdisziplinärer Sicht Ergebnisse eines Expertentreffens. Essen: Natur und Medizin; 2010.
10. Gesundheitsministerkonferenz. Beschluss der 87. Gesundheitsministerkonferenz am 26. und 27. Juni 2014 . TOP 7.3 Eckpunkte zur Überprüfung der für die Berufsausübung erforderlichen Deutschkenntnisse in den akademischen Heilberufen. Hamburg: Gesundheitsministerkonferenz; 2014. Zugänglich unter/available from: https://www.gmkonline.de/documents/TOP73BerichtP_OeffentI_Bereich.pdf

11. Marburger Bund. Deutschkenntnisse – Anforderungen in den Bundesländern für die Approbationserteilung Stand: Januar 2018. Berlin: Marburger Bund; 2018. Zugänglich unter/available from: https://www.marburger-bund.de/sites/default/files/files/2018-09/deutschkenntnisse-german-requirements-approbation.pdf
12. McNamara T. Item Response Theory and the validation of an ESP test for health professionals. Language Test. 1990;7(1):52-76. DOI: 10.1177/026553229000700105
13. Woodward-Kron R, Elder C. A Comparative Discourse Study of Simulated Clinical Roleplays in Two Assessment Contexts: Validating a Specific-Purpose Language Test. Language Test. 2016;33(2):251-270. DOI: 10.1177/0265532215607399
14. McNamara, T. Problematising content validity: the Occupational English Test (OET) as a measure of medical communication. Melbourne Papers. Language Test. 1997;6(1):19-43.
15. Moosbrugger H, Kelava A. Testtheorie und Fragebogenkonstruktion. Berlin: Springer; 2008. DOI: 10.1007/978-3-540-71635-8
16. Bachman L, Palmer A. Language testing in practice. Oxford: Oxford University Press; 2013.
17. Corkill D. Handbuch zur Entwicklung und Durchführung von Sprachtests. Frankfurt a.M.: Telc; 2012.
18. Nikendei C, Jünger J. OSCE – praktische Tipps zur Implementierung einer klinisch-praktischen Prüfung. GMS Z Med Ausbild. 2006;23(3):Doc47. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2006-23/zma000266.shtml
19. Brandes H. Überprüfung kommunikativer Fähigkeiten der Studierenden des Reformstudienganges Medizin der Charité Berlin mit Hilfe einer OSCE-Station. Berlin: Charité – Universitätsmedizin Berlin, Medizinischen Fakultät; 2006.
20. Bagnasco A, Tolotti A, Pagnucci N, Torre G, Timmins F, Aleo G, Sasso L. How to maintain equity and objectivity in assessing the communication skills in a large group of student nurses during a long examination session, using the Objective Structured Clinical Examination (OSCE). Nurse Educ Today. 2016;38:54-60. DOI: 10.1016/j.nedt.2015.11.034
21. Kiehl C, Simmenroth-Nayda A, Goerlich Y, Entwistle A, Schiekirka S, Ghadimi B, Raupach T, Koenig S. Standardized and quality-assured video-recorded examination in undergraduate education: informed consent prior to surgery. J Surg Res. 2014;191(1):64-73. DOI: 10.1016/j.jss.2014.01.048
22. Eckes T. Die Beurteilung sprachlicher Kompetenz auf dem Prüfstand. Fairness in der beurteilergestützten Leistungsmessung. In: Aguado K., Schramm K., Vollmer H, eds. Fremdsprachliches Handeln beobachten, messen, evaluieren Neue methodische Ansätze der Kompetenzforschung und der Videographie. Frankfurt a.M.: Lang; 2010. S.65-97.
23. Kecker G. Was macht eine gute Sprachprüfung aus? Qualitätssicherung beim TestDaF. In: Drumbl H, Kletschko D, Sorrentino D, Zanin R, eds. Lerngruppenspezifik in DaF, DaZ, DaM. Bozen: Bozen University Press; 2016. S.145-64.
24. Association of Language Testers in Europe (ALTE). Handreichungen für Testautoren. 2nd ed. Bochum: Association of Language Testers in Europe (ALTE); 2005. Zugänglich unter/available from: https://www.testdaf.de/fileadmin/Redakteur/Bilder/Aktuelles/2007/ALTE_Deutsche_HR_Vorwort.pdf
25. Fawcett T. An introduction to ROC analysis. Patt Recogn Lett. 2006;27(8):861-874. DOI: 10.1016/j.patrec.2005.10.010
26. Rice M, Harris G. Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. Law Human Behav. 2005;29(15):615-620. DOI: 10.1007/s10979-005-6832-7
27. Youden W. Index for rating diagnostic tests. Cancer. 1950;3(1):32-35. DOI: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Corresponding author:
Holger Lenz
Klinikum der Universität München, Institut für Didaktik und Ausbildungsforschung in der Medizin, Pettenkoferstr. 8A, D-80336 München, Germany
holger.lenz@med.uni-muenchen.de

Please cite as
Lenz H, Opitz A, Huber D, Jacobs F, Paik WG, Roche J, Fischer MR. Language Matters: Development of an Objective Structured Language Test for Foreign Physicians – Results of a Pilot Study in Germany. GMS J Med Educ. 2019;36(1):Doc2.
DOI: 10.3205/zma001210, URN: urn:nbn:de:0183-zma0012109

This article is freely available from
http://www.egms.de/en/journals/zma/2019-36/zma001210.shtml

Received: *2018-06-20*
Revised: *2018-12-05*
Accepted: *2018-12-19*
Published: *2019-02-15*

Sprache zählt: Entwicklung eines objektiven strukturierten Sprachtests für ausländische Ärztinnen und Ärzte – Ergebnisse einer Pilotstudie in Deutschland

Zusammenfassung

Zielsetzung: Entwicklung einer wissenschaftlich fundierten und standardisierten Fachsprachenprüfung für das Bundesland Bayern gemäß den Vorgaben der 87. Gesundheitsministerkonferenz (GMK). Der SAM – Sprachtest für ausländische Mediziner soll Teil des Approbationsverfahrens ausländischer Ärzte und Ärztinnen sein. In situativen Prüfungsstationen soll er fachsprachliche und kommunikative Kompetenzen auf C1-Niveau abprüfen.

Methodik: Für vier je zehnminütige Mini-Interviews wurden Fallvignetten ausgearbeitet, für die 40-minütige schriftliche Prüfungsstation, die aus zwei Teilaufgaben besteht, wurde ein Video einer Anamnese sowie kommentierte Laborergebnisse als Basis der Aufgabenstellungen erstellt. Fachsprachlichen Kompetenzen wurden anhand von Analysen wissenschaftlicher Literatur und empirischer Beispiele fixiert und als Items zu Bewertungsskalen für jede Teilstation zusammengefasst. In drei Simulationen wurden die Prüfungen per Video (SAM-Prüfungssoftware) aufgezeichnet und im Anschluss von Bewerterteams bewertet.

Ergebnisse: 19 Probanden nahmen an drei Simulationen teil. Eine Goldstandardsetzung konnte bei 18 von ihnen durchgeführt werden. Eine ROC-Analyse ergab einen AUC-Wert von .83, was die prognostische Qualität des SAM bestätigt. Die Reliabilität des SAM konnte nur für zehn Probanden berechnet werden. Die mit Cronbachs Alpha berechnete interne Konsistenz betrug .85. Die Bestehensgrenze wurde mithilfe des Youden-Index ermittelt. Für den SAM ergab sich dabei die Grenze von >60%.

Schlussfolgerung: Mit dem SAM wurde eine valide Fachsprachenprüfung mit hoher Test-Objektivität vorgelegt, die in authentischen Kommunikationssituationen und einem standardisierten Setting die Fachsprachenkenntnisse im geforderten C1-Niveau abprüft. Mit weiteren Erprobungen und einer größeren Stichprobe kann der SAM weiter validiert und eine höhere Test-Reliabilität sichergestellt werden.

Schlüsselwörter: Prüfung, Fachsprache, ausländische Ärzte

Holger Lenz[1]
Ansgar Opitz[2]
Dana Huber[3]
Fabian Jacobs[1]
Wolfgang Gang Paik[4]
Jörg Roche[5]
Martin R. Fischer[1]

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

2 LMU München, Lehrstuhl für Empirische Pädagogik und Pädagogische Psychologie, München, Deutschland

3 LMU München, (ehem.) Institut für Deutsch als Fremdsprache, München, Deutschland

4 LMU München, Medizinstudierender, München, Deutschland

5 LMU München, Institut für Deutsch als Fremdsprache, München, Deutschland

1. Einleitung

„Wer nur die leicht steigenden Arztzahlen betrachtet, verschließt die Augen vor der ganzen Wahrheit. Tatsächlich öffnet sich die Schere zwischen Behandlungsbedarf und Behandlungskapazitäten immer weiter.“ [1]. So kommentierte der Präsident der Bundesärztekammer Frank Ulrich Montgomery die bundesweite Ärztestatistik für das Jahr 2016. Längst sind „Ärztemangel“ und „Fachkräftemangel“ fest etablierte Schlagwörter im gesundheitspolitischen Diskurs [2]. Die Versorgungslücke schließen immer mehr Ärztinnen und Ärzte aus dem Ausland, deren Zahl sich in den letzten fünf Jahren fast verdoppelt und 2016 mit insgesamt 41.658 in Deutschland ein neues Rekordhoch erreicht hat [3].

Beim Integrationsprozess in den Berufsalltag sind diese aber mit fachbezogenen, administrativen und kulturellen Herausforderungen konfrontiert, die sich immer auch sprachlich manifestieren. Mangelnde oder mangelhafte Kompetenzen führen oft zu sinkender Behandlungsqualität, geringer Patientenzufriedenheit und interkollegialen Konflikten und gefährden somit erheblich die Patientensicherheit. Im Extremfall entscheidet das Scheitern von Kommunikation sogar über Leben und Tod [4], [5], [6], [7], [8]. Kompetente Kommunikation, die Missverständ-

nisse ausräumt und verhindert, ist ein vitales Element ärztlicher Praxis [9].

Daher beschloss die 87. Gesundheitsministerkonferenz (GMK) 2014 die bundesweite Einführung einer Fachsprachenprüfung (FSP) für Berufstätige in verkammerten akademischen Heilberufen unter Vorgabe grundlegender Mindestanforderung. Die Anforderungen schließen ein simuliertes Berufsangehöriger-Patienten-Gespräch, das Anfertigen eines in der ärztlichen Praxis vorkommenden Schriftstückes und ein Gespräch mit einem Angehörigen derselben Berufsgruppe ein. Für jeden Teil wurden 20 Minuten veranschlagt [10] (vgl. Abbildung 1). Bis dato gibt es keine gemeinsamen Standards für die testtheoretischen und methodischen Rahmenbedingungen der FSP. Die formellen Rahmenvorgaben der GMK beziehen sich vor allem auf das Sprachniveau C1 in einer fachsprachlichen Ausprägung. Mit diesen Vorgaben wurden zwar notwendige Rahmenbedingungen für höhere sprachliche Standards geschaffen; gleichzeitig liegt die Verantwortung, qualitativ hochwertige FSPen zu garantieren, bei den einzelnen Ländern. Anhand einer Übersicht des Marburger Bunds zeigt sich eindrücklich die Diversität in der Umsetzung der Sprachprüfung zwischen einzelnen Bundesländern [11]. Das Fehlen einer bundesweit einheitlichen FSP birgt wiederum die Gefahr des „Prüfungstourismus“, der darin besteht, dass sich ausländische Ärztinnen und Ärzte bevorzugt in Ländern zur Prüfung anmelden, in denen die Prüfung leichter zu bewältigen ist als in anderen Bundesländern. Der Freistaat Bayern, vertreten durch das Staatsministerium für Gesundheit und Pflege (StMGP), beauftragte 2016 daher ein interdisziplinäres Forscherteam der Ludwig-Maximilians-Universität (LMU) aus den Bereichen Medizin, Medizindidaktik, Deutsch als Fremdsprache (DaF) und Psychometrie mit der Entwicklung einer validen, reliablen, fairen, authentischen, objektiven und ökonomisch durchführbaren Sprachprüfung für ausländische Mediziner (SAM). Im englischsprachigen Raum gilt das australische Verfahren als führendes, da es auf ähnlichen wissenschaftlich-methodischen Standards aufbaut [12]. Eine Analyse dieses Verfahrens hat gezeigt, dass internationale Modelle zwar als Orientierungshilfe dienen können. Auch das australische Verfahren entspricht jedoch nicht allen Kriterien der wissenschaftlichen Testentwicklung [13], [14]. Eine eigenständige methodische Fundierung des SAM war deshalb unumgänglich. Der vorliegende Artikel skizziert die Konzeption und Pilotierung des SAM und stellt bisherige Ergebnisse vor.

2. Projektbeschreibung und Methodik

Unter Rücksichtnahme auf die im Eckpunktepapier der 87. GMK genannten Vorgaben entwickelte das SAM-Team ein Konzept, das v.a. die testtheoretischen Gütekriterien der Objektivität, Reliabilität, Validität und Authentizität erfüllen soll [10], [15], [16], [17].

2.1. Prüfungsaufbau

Der Aufbau des SAM ist in Abbildung 1 dargestellt. Für den Bereich Arzt-Patienten-Kommunikation wurde das Führen eines *Anamnesegesprächs*, sowie das Führen eines vorbereitenden *Aufklärungsgesprächs über eine Operation* (OPV) gewählt. Beim *Anamnesegespräch* muss der Prüfling die für eine Anamnese notwendigen Informationen vom Patienten einholen, ihm Raum zum Berichten über Beschwerden einräumen und eine respektvolle Gesprächsatmosphäre schaffen. Gleichzeitig wird die rezeptive Sprachkompetenz geprüft. Im Teilbereich OPV liegt der Fokus auf der Informationsvermittlung. Der Arzt soll dem Patienten den Ablauf einer bevorstehenden Operation, die Risiken des Eingriffs, sowie postoperative Verhaltensmaßnahmen vermitteln. Das Augenmerk liegt hier auf der Verwendung von Laiensprache (allgemeinsprachliche Ausdrücke statt medizinischer Fachbegriffe), dem Rückversichern, dass der Patient alle Informationen verstanden hat, sowie dem verbalen und nonverbalen Ausdruck von Empathie bei Bedenken und Fragen.

Als prototypische Kommunikationssituation für professionelle Interaktion wurde die *Patientenvorstellung* – der Stationsarzt berichtet dem Oberarzt – ermittelt. Hier soll der Geprüfte in einer simulierten Patientenvorstellung unter Einsatz berufssprachlicher Begriffe und Redewendungen das Kommunizieren unter Kollegen (hier: Oberarzt) unter Beweis stellen. Sowohl die Informationsweitergabe als auch Rückfragen sollen knapp und präzise formuliert werden.

Im Gegensatz zu anderen Fachsprachentests in Deutschland schließt der SAM auch die per GMK geforderte Überprüfung fachsprachlicher Kompetenzen zwischen Ärzten und Angehörigen anderer Heilberufe ein [10]. Als typische Kommunikationssituation wurde hierfür das Anweisungsgespräch *mit einem Krankenpfleger/einer -pflegerin* gewählt. Im Arzt-Pfleger-Gespräch werden klar verständliche Weisungen an einen Pfleger weitergegeben. Dies soll ebenfalls unter Verwendung berufssprachlicher Begriffe und Redewendungen in respektvoller Gesprächsatmosphäre geschehen.

Für den schriftlichen Teilbereich des Tests ergab eine Korpusanalyse von 200 Arztbriefen aus Chirurgie und Innerer Medizin am Klinikum der LMU hinsichtlich Struktur und sprachlicher Gestaltung, dass Arztbriefe in der Regel aus vier typischen Strukturelementen bestehen. Von diesen wurden zwei – *Anamnese mit Aufnahmegrund* und *Verlauf und Procedere* – aufgrund der hohen sprachlichen Anforderung in den schriftlichen Teil des SAM übernommen. Der schriftliche Teil prüft die Rezeptionsfähigkeit und Verarbeitung sprachlichen Inputs, sowie die schriftsprachliche Ausdrucksfähigkeit des Prüflings.

Für die Fallvignetten wurden Fälle aus den Fachbereichen „Allgemeinmedizin“, „Innere Medizin“ und „Chirurgie“ gewählt. Diese Bereiche decken sich weitgehend mit den Inhalten der Kenntnisprüfung, die ausländische Ärztinnen und Ärzte aus Drittstaaten (nicht EU) nach erfolgreichem Bestehen der Fachsprachenprüfung ablegen müssen,

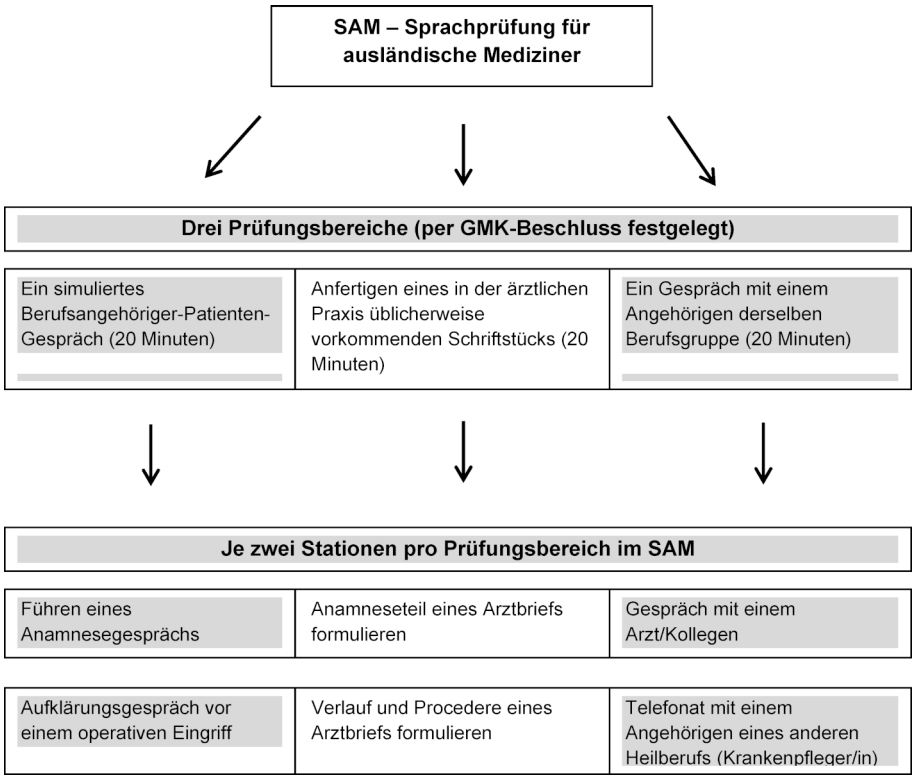

Abbildung 1: Schematischer Aufbau der Sprachprüfung für ausländische Mediziner unter Berücksichtigung der per GMK festgelegten Mindestanforderungen

um ihr medizinisch-fachliches Wissen auf Niveau des 3. Staatsexamen nachzuweisen, bevor sie die Approbation erhalten.
Unabhängig von der persönlichen fachlichen Spezialisierung der Prüflinge kann daher eine Konzentration auf diese Fachgebiete als gerechtfertigt angesehen werden. Um eine Fokussierung der Prüfung auf fachspezifische Inhalte zu vermeiden, wurden die Fallszenarien so allgemein wie möglich gehalten. Z.B. behandelt das Aufklärungsgespräch häufige chirurgische Eingriffe wie die Operation an der Schilddrüse oder die Tonsillektomie.

2.2. Prüfungsformat

Um den realen (authentischen) Anforderungen der Berufspraxis gerecht zu werden und damit gleichzeitig vergleichbare Bedingungen zum Nachweis medizinischer Kompetenzen im Studium anzusetzen, wurde das *OSCE*-Format (Objective Structured Clinical Examination) für den SAM gewählt. Nach Miller bieten OSCE-Prüfungen die Möglichkeit, Wissen nicht nur zu reproduzieren, sondern Gelerntes in kontextuell-situativer Praxis zu *zeigen* [17]. Aus medizindidaktischer Sicht haben sich OSCEs als reliables und valides Instrument zur Prüfung klinisch-praktischer Fähigkeiten international etabliert [18]. Brandes und Bagnasco et al. haben zudem gezeigt, dass sich OSCEs auch als methodisches Setting für die Messung kommunikativer Kompetenzen in kulturellen und professionellen Kontexten eignen [19], [20].
Analog zum OSCE-Konzept kurzer Prüfungsstationen von fünf bis zehn Minuten sieht der SAM jeweils zwei Stationen à zehn Minuten für jeden der zwei mündlichen Prüfungsbereiche vor (siehe Abbildung 1). Dies führt zu einer erhöhten Reliabilität, da das Verhalten des Prüfungsteilnehmers dadurch insgesamt viermal in unterschiedlichen Kontexten beobachtet werden kann. Zehnminütige Stationen stellen zudem eine realistische Abbildung der zeitlichen Ressourcen im Arbeitsalltag von Ärztinnen und Ärzten dar, was sich wiederum positiv auf das Authentizitätskriterium auswirkt.

2.3. Abhängigkeit der Prüfungsteile

Bestehende Fachsprachenprüfungen testen meist einen einzigen Fall über alle vorgegebenen Prüfungsbereiche hinweg. Aus psychometrischer Sicht ist dieses Konzept problematisch: kommt *ein* Prüfungsfall über die gesamte Prüfung zum Einsatz, entsteht dadurch ein Abhängigkeitsverhältnis zwischen den Bewertungskriterien der Prüfungsbereiche. Die Leistung in einem Bereich hängt dann nicht mehr ausschließlich von der Kompetenz in diesem Bereich ab, sondern auch von der Leistung in bereits durchlaufenen Testabschnitten [15].
Zugleich führt das Ein-Fall-Szenario zu einer deutlichen Reduktion der Fairness: bekommt der Teilnehmer z. B. zufällig einen Fall aus einem Fachgebiet, mit dem er oder sie durch bisherige ärztliche Tätigkeit besonders vertraut ist, führt dies automatisch zu einer besseren Prüfungsleistung und umgekehrt. Schließlich erleichtert das Modell voneinander unabhängiger Testbereiche den durch Prüfungsverschleiß bedingten Austausch älterer Fallszenarien: besteht ein Test aus mehreren Fällen, ist es möglich, die Schwierigkeit eines neu eingeführten Falls mit den bestehenden Fällen mit bekannter Schwierigkeit zu ver-

gleichen; besteht ein Test dagegen nur aus einem Fall, bedeutet der Austausch eines Falls automatisch den Austausch des gesamten Tests. Somit ist keine vergleichbare Einschätzung der Schwierigkeit des neuen Falls möglich. Für den SAM liegen daher unterschiedliche Fälle pro Prüfungsbereich zugrunde.

2.4. Durchführung und Bewertung

Jede Sprachprüfung, die die rezeptive und produktive Sprachleistung der Teilnehmer testen soll, muss Kommunikationssituationen schaffen, die so realitätsnah wie möglich (authentisch) und so wiederholbar wie möglich (objektiv und fair) sind. Damit wird gewährleistet, dass alle Teilnehmer im gleichen kommunikativen Kontext getestet werden. Um solch standardisierte Kommunikationssituationen zu schaffen, werden im SAM ausgebildete Schauspieler für die Rollen „Patient“ und „Pflegekraft“ eingesetzt. Die Rolle des vorgesetzten Arztes im Arzt-Arzt-Gespräch übernimmt ein Arzt.

Sowohl Schauspielpatienten als auch Arzt wurden in mehrstündigen Einheiten trainiert und geschult. Hauptaugenmerk der Schulungen lag dabei auf der Vereinheitlichung der Prüfungssituation (Objektivität, Fairness) und dem Evozieren fallspezifischer Sprachhandlungen. Pro mündliche Station wurde ein Skript für die Schauspielpatienten mit detaillierten Gesprächsanleitungen und Zusatzfragen erarbeitet.

Bereits bestehende Fachsprachprüfungen in anderen Bundesländern bewerten die Leistung der geprüften Person synchron: mehrere Prüfer sitzen mit im Raum und bewerten die Leistung des Prüflings, meist auf Grundlage vorgefertigter Bewertungsbögen. Synchrone Bewertungen mündlicher Prüfungsleistungen sind jedoch in verschiedener Hinsicht problematisch: das Gesagte ist flüchtig und kann nicht wiederholt werden; es wird zudem nur ‘aus der Situation’ bewertet und das oft von einer an der Kommunikationssituation beteiligten Person.

Eine asynchrone Bewertung mit unabhängigen Bewertern hingegen, die nur den mündlichen Text bewerten, ermöglicht das wiederholte, unabhängige, standardisierte Anhören der Prüfungsleistung und steigert somit die Auswertungsobjektivität. Im SAM werden die mündlichen Teilbereiche daher per Video aufgezeichnet. Diese VOSCE (*Video-Recorded Objective Structured Clinical Examination*) genannte Prüfungs- und Bewertungsform wurde bereits als durchführbare, reliable und valide Methode zur Bewertung kommunikativer Fähigkeiten in anderen medizinischen Kontexten erfolgreich erprobt [21], [22], [23]. Da Speicherung und Zugriff auf aufgezeichnete Prüfungsleistungen aus datenschutztechnischen Gründen oft problematisch ist, wurde hierfür eigens ein Computerprogramm entwickelt, das die Prüfungsleistungen über eine an einem Laptop angeschlossene Kamera aufzeichnet, diese auf einem geschützten Server pseudonymisiert speichert und dem Bewerterteam schließlich zu einem späteren Zeitpunkt sicheren Zugang zu den Dateien gewährt.

Das Bewerterteam besteht dabei aus einer Ärztin oder einem Arzt und einem Sprachwissenschaftler mit testmethodischem Fachwissen zu Deutsch als Fremdsprache. Für die Bewertung wurde pro Prüfungsstation (Anamnese, OPV, etc.) eine eigene Skala entwickelt. Die Bewerter wählen bei jedem Item eine von drei Antwortmöglichkeiten: „Trifft eher zu“, „Trifft eher nicht zu“ und „Uneindeutig“. Die Antwortoption „Trifft eher zu“ wird mit einem Punkt bewertet, die Option „Trifft eher nicht zu“ mit 0 Punkten und die Option „Uneindeutig“ mit 0,5 Punkten. Die zu bewertenden Items sind bezüglich der fachsprachentypischen Struktur, der sprachlichen Gestaltung und des kommunikativen Verhaltens, sowie der globalen Einschätzung des gesamten Gesprächs gruppiert. Pro Teilstation wurden zwischen 11 und 17 Items erstellt, was einer Gesamtzahl von 83 Items für den SAM insgesamt entspricht (siehe Tabelle 1). Eine Beispielskala für die Station Anamnesegespräch findet sich in Anhang 1.

Ein der Skala angehängtes Beiblatt erklärt die Intention und die Verwendung der Items im Bewertungsprozess und gibt fallspezifische Beispiele. Dies entspricht den Forderungen der Association of Language Testers in Europe (ALTE) zur Bewertung von Sprachtests [24] und steigert die Wahrscheinlichkeit einer einheitlichen Bewertung. Zusätzlich erhielt das Bewerterteam direkt vor der ersten Bewertung eine ca. einstündige Schulung durch die Testautoren, in denen das Bewertungsverfahren und die Skalen erklärt und Fragen beantwortet wurden.

Die Experten müssen einstimmig über das Bestehen oder Nicht-Bestehen eines Prüflings entscheiden, wobei die Bewertung zunächst getrennt erfolgt. Nach getrennter Bewertung vergleichen die Bewerter das Ergebnis und einigen sich bei abweichender Bewertung auf einen Wert. Die kumulative Leistung in den sechs Teilbereichen entscheidet schließlich über Bestehen oder Nichtbestehen.

3. Pilotierung des SAM

3.1. Durchführung

In der Pilotierungsphase wurde der Test in drei Simulationen erprobt. Insgesamt nahmen an den drei Simulationen 19 Prüflinge teil. Testteilnehmer waren dabei entweder ausländische Medizinstudierende der LMU (n=10), oder ausländische Ärztinnen und Ärzte, die noch keine Approbation in Deutschland haben (n=9). Mit Hilfe der Simulationen konnte einerseits die Durchführbarkeit des Tests überprüft werden. Andererseits wurden die Ergebnisse genutzt, um die Beobachterübereinstimmung, Reliabilität und prognostische Güte des SAM zu bestimmen, sowie, um die Bestehensgrenze festzulegen.

Um die prognostische Güte zu überprüfen und die Bestehensgrenze festzulegen, wurde ein sogenannter Goldstandard ermittelt: neben der regulären Beurteilung der Prüfungsleistungen von einem Bewerterteam, kam ein Expertenteam mit langjähriger Erfahrung in der Bewertung kommunikativer Prüfungsleistungen aus den Bereichen Medizin und Deutsch als Fremdsprache zum Einsatz. Diese Experten beurteilten auf globaler Ebene, ob die Prüflinge mindestens das C1-Niveau erreicht haben. Der

Tabelle 1: Reliabilitäten der Teilstationen des SAM

Teilstation (in Klammern Anzahl der Items und Personen, mit denen der Wert berechnet werden konnte)	Reliabilität (Cronbachs alpha)
Arztbrief - Anamnese (17 Items; N = 18)	.903
Arztbrief - Verlauf & Procedere (13 Items; N = 18)	.929
Patientengespräch - Anamnese (16 Items; N = 14)	.594
Patientengespräch - OP-Vorbereitung (12 Items; N = 14)	.224
Pflegeanweisung (12 Items; N = 13)	.675
Arzt-Arzt-Interaktion (11 Items; N = 15)	.911

Vergleich der regulär bewerteten Items mit diesem globalen Expertenurteil, das den sogenannten Goldstandard darstellt, erlaubt es die Qualität des SAM zu beurteilen und eine Bestehensgrenze festzulegen.

3.2. Ergebnisse

Die Übereinstimmung des Bewerterteams bei der Beurteilung der 83 Items lässt sich am besten mit Cohens Kappa ermitteln. Dies gibt an, inwieweit die Übereinstimmung der beiden Bewerter bzgl. der abgegebenen Bewertungen im Vergleich mit zufällig generierten Bewertungen höher ausfällt. Mögliche Werte liegen zwischen 0 und 1. Die so erfasste Übereinstimmung ließ sich durch die durchgeführten Schulungen von .49 auf .72 steigern. Die prozentuale Übereinstimmung lag am Ende der Pilotierungsphase bei 88% (zu Beginn: 80%).

Aufgrund fehlender Daten und leicht unterschiedlicher Itemzusammenstellungen zwischen der ersten und den weiteren beiden Erprobungen konnte die Reliabilität der Gesamtskala des SAM nur für zehn Prüflinge anhand 81 Items berechnet werden. Für die Menge dieser 81 Items betrug die mit Cronbachs alpha berechnete interne Konsistenz .85. Die Reliabilitäten der Teilstationen (für die mehr Daten vorhanden sind) können Tabelle 1 entnommen werden.

Im Durchschnitt erfüllten die Prüflinge M=55% (SD=20%) der 83 Items der sechs Skalen. Die Goldstandardsetzung konnte bei 18 Prüflingen durchgeführt werden. Fünf wurden dabei als fachsprachlich qualifiziert (auf dem C1-Niveau) eingestuft. Die Leistung dieser fünf Prüflinge im SAM lag dabei mit M=69% (SD=19%) über der Leistung derer, die laut Goldstandard nicht als fachsprachlich qualifiziert gelten (M=46%, SD=14%). Um die prognostische Qualität des SAM (in Bezug auf die Eignung der Prüflinge) genauer zu untersuchen, wurde eine sogenannte ROC-Analyse (receiver operating characteristic) durchgeführt [25]. Diese ermittelt, inwieweit die Prüfungsleistung in einem Test mit der „wirklichen“ Leistung der Prüflinge (die durch den Goldstandard abgebildet wird) übereinstimmt. Die globale Güte des Tests kann dabei mit Hilfe des AUC-Werts (area under the curve) quantifiziert werden. Der AUC-Wert kann zwischen 0 und 1 liegen. Eine AUC von .5 bedeutet, dass der Test nicht besser als der Zufall zwischen geeigneten und ungeeigneten Prüflingen unterscheidet. Eine AUC von 1 bedeutet, dass alle Prüflinge korrekt eingeschätzt werden. Für den SAM ergab sich eine AUC von .83, was nach gängigen Messverfahren einer großen Effektstärke entspricht und damit die prognostische Qualität des SAM unterstreicht [26].

Mit Hilfe der ROC-Analyse wurde zudem die Bestehensgrenze ermittelt. Dabei wurde der Youden-Index herangezogen [27]. Dieser Index kombiniert die Sensitivität (die Rate der qualifizierten Prüflinge, die vom Test korrekt erkannt werden) und Spezifität (die Rate der unqualifizierten Prüflinge, die vom Test korrekt erkannt werden) des Tests zu einem einzelnen Wert. Höhere Werte sind dabei wünschenswert. Eine Bestehensgrenze von >50% erzielte dabei den Wert .49. An dieser Grenze betragen die Sensitivität .80 und die Spezifität .69. Der PPV (positive predictive value; die Wahrscheinlichkeit, dass ein Prüfling wirklich qualifiziert ist, wenn die Bestehensgrenze erreicht wird) liegt an dieser Schwelle bei .50 und der NPV (negative predictive value; die Wahrscheinlichkeit, dass ein Prüfling wirklich nicht qualifiziert ist, wenn die Bestehensgrenze nicht erreicht wird) liegt bei .90.

Eine Bestehensgrenze von >60% erreicht einen Youden-Index von .52. Zwar sinkt die Sensitivität auf .60, aber die Spezifität steigt dafür auf .92. Der PPV beträgt .75 und der NPV .86. Wenn man den Youden-Index als Kriterium heranzieht und davon ausgeht, dass es die höchste Priorität eines Tests ist, der über die Zulassung zur Arbeit als Arzt entscheidet, möglichen Schaden von der Bevölkerung abzuwenden, so sollte die konservative Grenze von >60% herangezogen werden. „Konservativ“ bedeutet in diesem Zusammenhang, dass ein Prüfling im Grenzbereich eher als unqualifiziert eingestuft wird. Die Daten der Erprobung erlauben sogar die Möglichkeit, die Grenze auf >70% zu legen. Ohne Sensitivitätsverlust würde dabei die Spezifität auf 1 steigen. Da die Verteilung der Daten allerdings nahelegt, dass die Sensitivität bei einer größeren Datenmenge bei einer solchen Grenze abfallen würde,

und da die Spezifität mit .92 bereits sehr hoch ist bei einer Grenze von >60%, wird zu einer Grenze von >60% für den SAM geraten. Tabelle 2 bietet eine Übersicht der wichtigsten Ergebnisse.

Tabelle 2: Übersicht der wichtigsten Test-Statistiken des SAM

Statistik	Wert
Reliabilität Gesamttest[a]	.85
Reliabilität Teilstationen[a]	.22 - .93
Interrater-Objektivität[b]	.72
AUC	.83
Sensitivität[c]	.60
Spezifität[c]	.92
PPV[c]	.75
NPV[c]	.86

Anmerkungen:
a) gemessen als Cronbachs alpha;
b) gemessen als Cohens Kappa;
c) Wert gilt an der empfohlen Bestehensgrenze von > 60%

4. Diskussion und Schlussfolgerung

Besonders in den Bereichen der Test-Fairness, Authentizität und Objektivität konnten gute Ergebnisse erzielt werden. In diesem Zusammenhang ist die Wichtigkeit der Schulung der an der Kommunikationssituation beteiligten Schauspielpatienten nochmals zu betonen. Erst das korrekte Verhalten der Schauspielpatienten dem jeweiligen Prüfungsteilnehmer gegenüber gewährleistet eine konstante Testumgebung. Die dadurch erhöhte Test-Objektivität wirkt sich ihrerseits positiv auf die Reliabilität und Validität aus. So könnte umgekehrt auch die geringe Reliabilität der Teilstation OP-Vorbereitung eventuell durch nicht vorgesehene Hilfestellungen der Schauspielpatienten, die teilweise bei der Erprobung beobachtet wurden, erklärt werden. Es könnte sein, dass die Schauspielpatienten (die keinen medizinischen Hintergrund besitzen) aus Mitgefühl mit schwächeren Prüfungsteilnehmern diesen Stichworte geben. Dies würde die systematische Varianz der Ergebnisse und damit die Reliabilität reduzieren. Diese und andere im Rahmen der vorliegenden Arbeit gesammelten Daten zu den speziellen Anforderungen und Schwierigkeiten, denen die Schauspielpatienten und Prüfungsteilnehmer im Rahmen der Prüfungssituation begegnen, können dabei als erste Grundlage für die Entwicklung einer standardisierten, wissenschaftlich gesicherten Schulungsmethodik dienen.

Eine weitere Stärke des SAM bietet der Ansatz der asynchronen Bewertung der Prüfungsleistung. Prüfer, die die Kommunikationssituation selbst miterleben oder sogar selbst daran beteiligt sind, erhöhen die Gefahr der Verzerrung (*Bias*) der Leistungsbewertung. Das im SAM verfolgte Modell der asynchronen Bewertung trägt zu einer fairen und objektiven Bewertung aller Teilnehmer bei und reduziert somit das Potential rechtlicher Beschwerden seitens der Prüflinge.

Die Validitätswerte des SAM basierend auf der ROC-Analyse der Pilotierungsdaten sind vielversprechend. Dies gilt besonders, wenn man bedenkt, dass die Rate geeigneter Teilnehmer laut Goldstandard gering war, was die Erkennung der geeigneten Kandidaten erschwert. Bei der Bewertung der Ergebnisse muss dabei bedacht werden, dass es sich bei mehr als der Hälfte der Probanden um ausländische Studierende handelte. Da Studierende im Vergleich zu bereits erfahrenen Ärzten insgesamt geringere Kenntnisse mitbringen, kann dies zu einer Verzerrung des Gesamtbilds der Prüfungsleistungen aller Teilnehmer beitragen. In einer Stichprobe erfahrener Ärzte sollte die Rate geeigneter Kandidaten höher liegen. Des Weiteren muss bedacht werden, dass die Unsicherheit aller Kennwerte aufgrund der relativ kleinen Erprobungs-Stichprobe noch hoch ist. Der Test sollte daher dringend systematisch validiert werden. Beispielsweise standen der insgesamt guten Skalenvalidität während der bisherigen Testsimulation und -auswertung unzureichend abgesicherte Reliabilitäten zweier Einzelskalen gegenüber (OP-Vorbereitung und Pflegeanweisungen). Zukünftige Simulationen, die die Schwäche der genannten Einzelskalen verringern und die psychometrische Qualität aller Skalen verbessern wollen, profitieren folglich besonders von Stichproben, deren Umfang größer und deren Verhältnis zwischen Prüfungsteilnehmern auf dem C1-Niveau und solchen unterhalb des C1-Niveaus ausgeglichener wäre. Die genauere Bestimmung der Test-Reliabilität wäre eine weitere natürliche Folge einer erweiterten Datenmenge.

Eine weitere Schwäche des SAM ist, dass zu Beginn einige Investitionen (Aufnahmesoftware und Schulung der Schauspielpatienten und Bewerterteams) in die Prüfungslogistik getätigt werden müssen. Im Langzeitbetrieb sollten die Stärken des SAM diesen Nachteil allerdings mehr als ausgleichen.

Weiterer Handlungsbedarf liegt in der Verteilung der Itemanzahl der Bewertungsskalen. Die Itemanzahl der verschiedenen Teilstationen schwankt zwischen 11 und 17. Um die gleiche Gewichtung aller Teilstationen sicherzustellen, wird eine nachträgliche Anpassung empfohlen um die Itemanzahl nicht künstlich erhöhen oder reduzieren zu müssen. Bevor der Summenwert für den Test berechnet wird, müsste dabei die Punktzahl der sechs Teilstationen durch eine Multiplikation mit unterschiedlichen Faktoren so gewichtet werden, dass jeweils $^1/_6$ der maximal möglichen Gesamtpunktzahl in jeder Teilstation erworben werden kann.

5. Ausblick

Beim SAM handelt es sich um den ersten und bis dato einzigen bundesweiten Ansatz einer wissenschaftlich fundierten Fachsprachprüfung. Zusätzlich zu den per GMK-Beschluss vorgegebenen Rahmenbedingungen wurden testtheoretische Gütekriterien wie Objektivität, Reliabilität, Validität, Authentizität, Fairness und Ökonomie als Leitprinzipien in der Testentwicklung verfolgt. Als bisher einzige Fachsprachenprüfung in Deutschland bezieht der SAM die interprofessionelle Kommunikation mit ein. Neben der bisher erprobten Kommunikation zwischen einem Arzt und einer Pflegedienstleitung ist die Ausweitung auf weitere Situationen der interprofessionellen Kommunikation denkbar. Im Hinblick auf einen wissenschaftlich und damit auch rechtlich soliden Test muss der SAM jedoch in weiteren Simulationen erprobt und getestet werden.
Des Weiteren wird empfohlen, testmethodische Kennwerte anderer Testverfahren mit denen des SAM zu vergleichen. Nur so kann das Ziel eines bundesweit einheitlichen Fachsprachentests, der ausländische Ärzte reliabel auf C1 Niveau prüft und somit die Patientensicherheit gewährleistet, letztendlich erreicht werden. Derzeit ist eine weitere Validierung der in Bayern eingesetzten Verfahren mittels einer Vergleichsstudie geplant. Es ist das erklärte Ziel der Testentwickler, den Sprachtest mittelfristig als Beitrag zu einer nachhaltigen Verbesserung gegenwärtiger Testverfahren zum Einsatz zu bringen, nicht nur in Bayern.

Danksagung

Für die nachhaltige Unterstützung des Projekts bedanken wir uns ferner bei Prof. Dr. Matthias Siebeck, Klinik für Allgemeine, Viszeral-, Transplantations-, Gefäß- und Thoraxchirurgie der LMU München.

Förderung

Wir bedanken uns beim Bayerischen Staatsministerium für Gesundheit und Pflege (StMGP) für die Unterstützung des Projekts unter dem Förderkennzeichen G32g-G8517.1-2015/5-91.

Interessenkonflikt

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

Anhänge

Verfügbar unter
http://www.egms.de/en/journals/zma/2019-36/zma001210.shtml

1. Anhang_1.pdf (147 KB)
   Sprachprüfung für ausländische Mediziner (SAM)

Literatur

1. Bundesärztekammer. Die Schere zwischen Behandlungsbedarf und Behandlungskapazitäten öffnet sich. Berlin: Bundesärztekammer; 2017. Zugänglich unter/available from: http://www.bundesaerztekammer.de/presse/pressemitteilungen/news-detail/die-schere-zwischen-behandlungsbedarf-und-behandlungskapazitaeten-oeffnet-sich/
2. Bundesärztekammer, Kassenärztliche Bundesvereinigung. Dem deutschen Gesundheitswesen gehen die Ärzte aus! Studie zur Altersstruktur- und Arztzahlentwicklung. 5th ed. Berlin: Bundesärztekammer und Kassenärztliche Bundesvereinigung; 2010. Zugänglich unter/available from: http://www.kbv.de/media/sp/Arztzahlstudie_2010.pdf
3. Bundesärztekammer. Ärztestatistik zum 31. Dezember 2016. Berlin: Bundesärztekammer; 2017. Zugänglich unter/available from: http://www.bundesaerztekammer.de/fileadmin/user_upload/downloads/pdf-Ordner/Statistik2016/Stat16AbbTab.pdf
4. Wichmann R. Weitere Zehn Jahre Warten hilft nicht. Praxisguide D Krankenhaus. 2015;(1):14-15.
5. Karimi P, Rudenko O. Am Anfang verstand ich null. Praxisguide D Krankenhaus. 2015;(1):20-21.
6. Arndt J. Sprachbarrieren im Krankenhaus – Wenn dem Arzt die Worte fehlen. Pneumologie. 2016;70(9):564-566. DOI: 10.1055/s-0042-114156
7. AG Leipzig. Aufklärung durch einen Arzt, der die deutsche Sprache nicht beherrscht. MedR. 2003;10:582-583.
8. Roche J. Zur Frage der Deutschkenntnisse. Sprache Beruf. 2014;7:316-318.
9. Schröder H. Theoretische Aspekte der Arzt-Patienten-Interaktion. In: Witt C, ed. Der gute Arzt aus interdisziplinärer Sicht Ergebnisse eines Expertentreffens. Essen: Natur und Medizin; 2010.
10. Gesundheitsministerkonferenz. Beschluss der 87. Gesundheitsministerkonferenz am 26. und 27. Juni 2014 . TOP 7.3 Eckpunkte zur Überprüfung der für die Berufsausübung erforderlichen Deutschkenntnisse in den akademischen Heilberufen. Hamburg: Gesundheitsministerkonferenz; 2014. Zugänglich unter/available from: https://www.gmkonline.de/documents/TOP73BerichtP_OeffentI_Bereich.pdf
11. Marburger Bund. Deutschkenntnisse – Anforderungen in den Bundesländern für die Approbationserteilung Stand: Januar 2018. Berlin: Marburger Bund; 2018. Zugänglich unter/available from: https://www.marburger-bund.de/sites/default/files/files/2018-09/deutschkenntnisse-german-requirements-approbation.pdf
12. McNamara T. Item Response Theory and the validation of an ESP test for health professionals. Language Test. 1990;7(1):52-76. DOI: 10.1177/026553229000700105
13. Woodward-Kron R, Elder C. A Comparative Discourse Study of Simulated Clinical Roleplays in Two Assessment Contexts: Validating a Specific-Purpose Language Test. Language Test. 2016;33(2):251-270. DOI: 10.1177/0265532215607399
14. McNamara, T. Problematising content validity: the Occupational English Test (OET) as a measure of medical communication. Melbourne Papers. Language Test. 1997;6(1):19-43.
15. Moosbrugger H, Kelava A. Testtheorie und Fragebogenkonstruktion. Berlin: Springer; 2008. DOI: 10.1007/978-3-540-71635-8
16. Bachman L, Palmer A. Language testing in practice. Oxford: Oxford University Press; 2013.
17. Corkill D. Handbuch zur Entwicklung und Durchführung von Sprachtests. Frankfurt a.M.: Telc; 2012.

18. Nikendei C, Jünger J. OSCE – praktische Tipps zur Implementierung einer klinisch-praktischen Prüfung. GMS Z Med Ausbild. 2006;23(3):Doc47. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2006-23/zma000266.shtml

19. Brandes H. Überprüfung kommunikativer Fähigkeiten der Studierenden des Reformstudienganges Medizin der Charité Berlin mit Hilfe einer OSCE-Station. Berlin: Charité – Universitätsmedizin Berlin, Medizinischen Fakultät; 2006.

20. Bagnasco A, Tolotti A, Pagnucci N, Torre G, Timmins F, Aleo G, Sasso L. How to maintain equity and objectivity in assessing the communication skills in a large group of student nurses during a long examination session, using the Objective Structured Clinical Examination (OSCE). Nurse Educ Today. 2016;38:54-60. DOI: 10.1016/j.nedt.2015.11.034

21. Kiehl C, Simmenroth-Nayda A, Goerlich Y, Entwistle A, Schiekirka S, Ghadimi B, Raupach T, Koenig S. Standardized and quality-assured video-recorded examination in undergraduate education: informed consent prior to surgery. J Surg Res. 2014;191(1):64-73. DOI: 10.1016/j.jss.2014.01.048

22. Eckes T. Die Beurteilung sprachlicher Kompetenz auf dem Prüfstand. Fairness in der beurteilergestützten Leistungsmessung. In: Aguado K., Schramm K., Vollmer H, eds. Fremdsprachliches Handeln beobachten, messen, evaluieren Neue methodische Ansätze der Kompetenzforschung und der Videographie. Frankfurt a.M.: Lang; 2010. S.65-97.

23. Kecker G. Was macht eine gute Sprachprüfung aus? Qualitätssicherung beim TestDaF. In: Drumbl H, Kletschko D, Sorrentino D, Zanin R, eds. Lerngruppenspezifik in DaF, DaZ, DaM. Bozen: Bozen University Press; 2016. S.145-64.

24. Association of Language Testers in Europe (ALTE). Handreichungen für Testautoren. 2nd ed. Bochum: Association of Language Testers in Europe (ALTE); 2005. Zugänglich unter/available from: https://www.testdaf.de/fileadmin/Redakteur/Bilder/Aktuelles/2007/ALTE_Deutsche_HR_Vorwort.pdf

25. Fawcett T. An introduction to ROC analysis. Patt Recogn Lett. 2006;27(8):861-874. DOI: 10.1016/j.patrec.2005.10.010

26. Rice M, Harris G. Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. Law Human Behav. 2005;29(15):615-620. DOI: 10.1007/s10979-005-6832-7

27. Youden W. Index for rating diagnostic tests. Cancer. 1950;3(1):32-35. DOI: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Korrespondenzadresse:
Holger Lenz
Klinikum der Universität München, Institut für Didaktik und Ausbildungsforschung in der Medizin, Pettenkoferstr. 8A, 80336 München, Deutschland
holger.lenz@med.uni-muenchen.de

Bitte zitieren als
Lenz H, Opitz A, Huber D, Jacobs F, Paik WG, Roche J, Fischer MR. Language Matters: Development of an Objective Structured Language Test for Foreign Physicians – Results of a Pilot Study in Germany. GMS J Med Educ. 2019;36(1):Doc2.
DOI: 10.3205/zma001210, URN: urn:nbn:de:0183-zma0012109

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/zma/2019-36/zma001210.shtml

Eingereicht: *20.06.2018*
Überarbeitet: *05.12.2018*
Angenommen: *19.12.2018*
Veröffentlicht: *15.02.2019*