

Measuring competency-relevant knowledge in the competency-oriented student progress test

Abstract

Background: Since 2013 a competency-oriented student progress test (SKPT) has been administered at a number of German medical schools. The questions are generated on the basis of a two-dimensional blueprint, on which one axis contains the five competency domains – communicative competence (CO), practical clinical competence (CP), theoretical clinical competence (CT), scientific competence (SC), and professional decision-making competence (PR) – that form part of the competency model of the National Competency-based Catalogue of Learning Objectives for Undergraduate Medicine (NKLM). The feedback for students is structured in part according to these domains. The aim of our study is to examine if

1. the results differentiated by competency domain show adequate measurement accuracy and
2. if the results for the different domains also contain different information.

Methods: The SKPTs for the years 2013 to 2017, taken by a total of 3,027 students, were examined. The measurement accuracy was determined using the coefficient g_{lb} (greatest bound to reliability) and the standard error of measurement; discriminant analysis of the principal components was carried out to demonstrate differentiation between the competency domains.

Results: The reliability of the competency domains was above 0.8 for all SKPTs; exceptions to this were seen in two of the tests for CO and PR that had a reliability of 0.7–0.8. The results for all of the individual competency domains differed in their informational content compared to the overall of the other domains; the same applies for all pairwise comparisons, with the exceptions of CP and CT.

Discussion: The SKPT feedback for students that is differentiated by competency domains basically fulfills the requirements for measurement reliability and distinctness. An improvement of the measurement quality for CO and PR and a better differentiation between CP and CT is desirable.

Keywords: progress test, reliability, discriminant validity

Andreas Möltner¹

Stefan Wagener¹

Mirka Burkert¹

1 Medical Faculty of
Heidelberg, Baden-
Württemberg Center of
Excellence for Assessment in
Medicine, Heidelberg,
Germany

1. Background

Competency-based medical education (CBME) has received special attention over the past 20 years in connection with curriculum development at medical schools and health policy [3], [26]. As a consequence, this must also be reflected in educational assessments [5]. This has led to a veritable flood of publications on competency-based testing (see the critical discussion of this in [11]), with a particular focus on practical and workplace-based methods to assess communication skills, professional competence, and so on. Less attention has been paid to assessments using traditional multiple-choice questions (MCQs) since, assuming that competencies are primarily action-based, less importance is placed on the simple measure-

ment of knowledge. Nonetheless, knowledge is an essential prerequisite for taking competent action and can be measured using competency-based tests of knowledge [23]. MCQs continue to offer substantial advantages to accomplish this. Due to the generally brief amount of time needed to answer a MCQ, it is possible to ask a large number of questions on a test that then, relative to the time needed, allows for distinctly better coverage and representation of the curricular content that is being tested than do open-ended essay formats, which are laborious to grade, or practical assessment formats. In addition, objectivity and a high reliability are easier to achieve than, for instance, with workplace-based assessments in which there is the risk that their seemingly

higher validity is rendered void by insufficient standardization.

As part of the BMBF-funded project *Medical Education Research – Lehrforschung im Netz BW* (MERLIN, <http://www.merlin-bw.de>), a competency-oriented student progress test (SKPT) was designed in 2013 by the Baden-Württemberg Center of Excellence for Assessment in Medicine (Kompetenzzentrum für Prüfungen in der Medizin Baden-Württemberg).

Progress tests are recognized and used in medical education to map learning progress over the course of medical study [28]. To do this, students of all semester levels (year of study) are given the same test. If there is sufficient equivalence (same difficulty level) between the tests administered in consecutive years, students are able to see their progress over the course of their studies and receive corresponding feedback. Progress tests are administered at different medical schools in Germany [18]. In general, these tests have two main functions: a) to give students ongoing feedback on their individual academic proficiency and b) to give medical schools the opportunity to monitor curricula, observe academic progress in different student cohorts and compare different curricula [24], [34].

The questions on the SKPT derive from a two-dimensional blueprint whose axes contain the eight subjects identified by the German medical licensing regulations (ÄAppO) and the individual competencies that are grouped into five competency domains in the National Competency-based Catalogue of Learning Objectives for Undergraduate Medical Education (NKLM) [4]. This blueprint was created by three inter-university and interdisciplinary expert groups who were tasked with grouping the domains in the NKLM in clusters referred to as “competency domains” and creating “subject groups” out of the subjects listed in the medical licensing regulations. These groups needed to reflect the best balance possible between preclinical and clinical subjects (see table 1 and table 2). A special aspect of the SKPT is that the questions are generated by students who are trained during multiple annual workshops. These students compose questions based on the competencies defined in the NKLM with reference to the subject groups in order to fill in the cells of the blueprint in table 2. The number of existing questions per cell is documented in an ongoing manner so that, especially near the end of the process, only the questions still needed to fill the empty cells need to be formulated. See [33] for a detailed description.

The progress test is administered each year in November/December (as of 2015 in cooperation with the Institut für Kommunikations- und Prüfungsforschung gGmbH). All medical students enrolled at universities where the SKPT is offered can take the test. In 2017 the SKPT was administered at a total of 16 medical schools: Dresden, Erlangen-Nürnberg, Freiburg, Gießen, Hannover, Heidelberg, Homburg, Krems (Austria), Leipzig, Magdeburg, Mannheim, Marburg, LMU München, Tübingen, Ulm, and Witten/Herdecke. Participation is voluntary, except at the

private Karl Landsteiner University in Krems, Austria, where participation is mandatory.

The test consists of 120 Type A MCQs (one correct response out of four to five possible responses) plus the additional option of “I don’t know” and ten situational judgment test questions (SJT) to measure social competencies [20]. An exception to this was the first progress test in 2013 that consisted of 144 Type A questions and no SJT.

The “I don’t know” option is frequently used in progress tests and other formative tests to enable test-takers to explicitly document their knowledge deficits and to avoid any guessing based on the possible response options [13], [22].

Example question from the Theoretical Clinical Competency domain (CT) on the 2017 SKPT

You are treating a 12-year-old child with pneumonia on the pediatric ward. The patient has a history of frequent respiratory complaints and infections. During an ultrasound examination you determine situs inversus, meaning that the organs are reversed from their normal positions. As part of a bronchoscopy you then send a biopsy for histological analysis.

Which diagnosis are you expecting based on the histology?

1. Defect in the cell-to-cell junctions
2. Defect in the kinocilia
3. Formation of pseudostratified epithelium
4. Absence of surface differentiations
5. Defect in the basement membrane
6. I don’t know

The correct answer is B. Additional examples are included in the attachment 1. All of the SKPT questions and their explanations can be found on the publicly accessible webpage <https://www.komp-pt.de/fragen-aus-dem-progresstest/>.

After taking the test all examinees receive feedback on their performance (number of points earned) which is differentiated according the subject groups and competency domains. This feedback is absolute (criteria-based), relative in comparison to the other participants at the same semester level (standards-based), and – if prior tests have been taken – longitudinal to show the gain in knowledge (progress) compared to earlier scores (ipsative).

The aim of this study is to ascertain

1. if the scores differentiated for each competency domain have sufficient measurement precision and
2. if the questions of the different competency domains constitute empirically distinct clusters.

Table 1: Blueprint axes for the SKPT: assignment of the individual competencies in the National Competency-based Catalogue of Learning Objectives in Undergraduate Medicine (NKLM) to the competency domains of the competency-oriented student progress test (SKPT) and assignment of the subjects contained in the German medical licensing regulations (ÄAppO) to the subject groups.

Competencies in the NKLM	Competency domains of the SKPT
Collaborator, Communicator, Health Advocate, Preventive healthcare	Communicative competence (CO)
Use of diagnostic procedures, Practical clinical skills, Emergency measures, Therapeutic approaches	Practical clinical competence (CP)
Scholar	Scholarly competence (SC)
Professional decision-maker, Ethics and Law, Manager	Professional decision-making competence (PR)
Principles of normal function and structure, Pathophysiolo-gical mechanisms, Therapeutic principles, Diagnostic procedures, Preventive measures	Theoretical clinical competence (CT)

Subjects in the ÄAppO	Subject groups
Anatomy, Biology, Forensic Medicine, (Imaging Procedures), Radiation Therapy and Radiation Protection, Physics, Pathology, Clinical Pathological Conference	I
Internal Medicine, Hygiene, Microbiology and Virology, Infectious Diseases and Immunology, Physiology	II
Surgery and Orthopedics, Urology, Otorhinolaryngology, Alternative Medicine, Ophthalmology, Anesthesiology, Emergency Medicine, Dermatology and Venerology	III
Medical Sociology, Clinical Environmental Medicine, Occupational Medicine, Social Medicine, Epidemiology, Medical Biometrics and Medical Informatics, Health Economy and Health System, Public Health, History, Theory and Ethics of Medicine, Medical Terminology	IV
Medical Psychology, Psychosomatic Medicine and Psychotherapy, Psychiatry and Psychotherapy, Neurology	V
General Practice, Gerontology and Geriatrics, Pain Management, Palliative Medicine, Prevention, Rehabilitation, Physical Medicine and Naturopathy	VI
Gynecology and Obstetrics, Pediatrics, Human Genetics	VII
Biochemistry/Molecular Biology, Chemistry, Clinical Chemistry and Lab Diagnostics, Clinical Pharmacology and Pharmacotherapy, Pharmacology and Toxicology	VIII

Table 2: Blueprint of the competency- oriented student progress test 2014-2017 (see tab. 1 for explanation of the competency domains and subject groups).

Competency domain	Subject groups								Total	Percent (approx.)
	I	II	III	IV	V	VI	VII	VIII		
CO	2	5	5	2	2	2	2	2	22	20.0%
CP	3	6	6	3	3	3	3	3	30	25.0%
SC	2	3	3	2	2	2	2	2	18	12.5%
PR	2	4	4	2	2	2	2	2	20	17.5%
CT	3	6	6	4	3	3	3	3	30	25.0%
Total	12	24	24	12	12	12	12	12	120	
Percent	10%	20%	20%	10%	10%	10%	10%	10%		

Both of these issues are especially important for the usefulness of the feedback given to students:

1. feedback is only beneficial if the graded performance has been reliably measured; and
2. differentiated feedback is only meaningful if the individual scores reflect different content and thus are not redundant.

In the terminology used by Campbell and Fiske [1], the second issue involves proving “discriminant validity” (also referred to as “discriminative validity”). The term “validity” has been the subject of intense discussion over the past 30 years. Many authors critically view the use of different terms to describe validity, e.g. predictive, convergent, discriminant; (a thorough description of this can be found in [17]). In the present study we use this term to refer to its “classic” definition: the questions assigned to competency domains form “scales” which measure the performance of different tasks (see the discussion in [6]). These “scales” are also supposed to be reflected in the responses of the progress test-takers. Basically, questions grouped together in the same domain should be answered similarly well (or badly).

2. Methods

2.1. Data collection

The SKPT was administered once annually between 2013 and 2017. In the first two years the SKPT was taken as a paper-and-pencil test at each participating medical school; as of 2015 it is available online.

Test announcement, student registration and conducting the progress test at each university are handled individually by each medical school. The overarching coordination is in the hands of the Baden-Württemberg Center of Excellence for Assessment in Medicine (Kompetenzzentrum für Prüfungen in der Medizin Baden-Württemberg), which is housed at the Medical Faculty of Heidelberg. For more details on the administration of the test, please see [33], [29], [30], [31], [32].

The SKPT is designed as a formative test that can be taken voluntarily. However, at the medical school in Krems, Austria, the SKPT is mandatory. Since it has been shown in various analyses (not presented here) that the Krems student group differs clearly from the voluntary participants at the other medical schools, the Austrian group has been excluded from the following analysis. A second exclusion criterion was the number of answered questions. The present study includes only those test-takers who answered at least 100 of the 120 questions, meaning that they marked one of the four or five possible answers or indicated that they did not know (see table 3). Only complete datasets have been included for the first two paper-based SKPTs from 2013 and 2014. Table 4 shows the number of participants in the analysis by year of study. The ten SJT questions from 2014 through 2017 did not fall within the scope of this study.

Prior to taking the SKPT, the students consented to the use of their pseudonymized data for the purpose of quality assurance and academic research.

2.2. Statistical analyses

2.2.1. Evaluation of the questions

The questions that were generated based on the blueprint were type-A MCQs with the additional option to respond with “I don’t know.” One point is assigned if the correct answer is chosen. For this analysis, incorrect responses, “I don’t know” and questions left blank were treated the same and assigned zero points (see [22] for alternative grading approaches in which “I don’t know” and incorrect responses are treated differently). After test administration, there was a second review (post-review) of the questions, which was based on student comments and the statistical analysis. If it was determined that more than one response was possible for a question, the test-takers received one point if they had selected one of the correct answers. Based on the post-reviews, we had to exclude between three to eight questions on the SKPTs for the years 2013 to 2017 for being flawed (see table 3).

2.2.2. Reliability and measurement precision

The greatest lower bound of reliability (glb) is used to estimate the reliability of the competency domains [7], [25]. This is also the optimal algebraic estimate of the reliability for non-homogenous scales (in this case, Cronbach’s α , as a measure of the internal consistency, yields an underestimation of reliability).

Reliability is a relative measure of measurement precision based on the participant population; the standard error of measurement serves as an absolute measure of measurement precision and is calculated from the reliability and the standard deviation of the scale values for the examinees [9].

When comparing students who are at the same semester level, reliability is relevant also as it refers to these particular subpopulations. In these cases, lower values are to be expected in comparison to the overall reliability since, assuming almost the same measurement error, the variance in the number of correctly answered questions among examinees at the same academic level is lower than the number for all examinees across all semester levels (see [33]).

2.2.3. Delineating the competency domains (“discriminant validity”)

Testing whether or not the different competency domains denote empirical differences becomes somewhat complex given the progress test’s design: the individual questions are not only assigned to one competency domain, but also to a subject or a subject group. As a consequence of this design, the constructs defined by the two axes of

Table 3: Format, number of items and number of test-takers for the SKPTs (excluding Krems medical school) in the present study. The datasets for examinees who answered at least 100 questions were included in the analysis. The number in parentheses in the last column indicates the number of examinees who answered all questions.

Progress test	Format	Number of schools	MC items/ Excluded post-review	MC items in the analysis	Examinees	Examinees in the analysis (all questions answered)
2013	Paper	8	144 (-4)	140	434	433 (433)*
2014	Paper	8	120 (-6)	114	302	302 (302)
2015	Online	11	120 (-3)	117	1125	845 (832)
2016	Online	14	120 (-4)	116	1246	713 (694)
2017	Online	15	120 (-8)	112	1268	735 (726)

*One examinee with information missing on year of study was excluded.

Table 4: Number of participants in the analysis by year of study

Progress test	Year of study						Total
	1	2	3	4	5	6	
2013	43	38	66	114	155	17	433
2014	82	27	54	43	81	15	302
2015	190	94	123	194	191	53	845
2016	118	84	113	177	140	81	713
2017	165	98	114	143	142	73	735
Total	598	341	470	671	709	239	3028

the blueprint already overlap each other (for a detailed explanation of construct overlap, see [35], among others). For this reason, we cannot assume that the competency domains are directly apparent in the data when applying methods of factor or cluster analysis (compare the factor analysis of the validity and reliability of competency constructs in [12]).

Therefore, we chose Fisher's linear discriminant analysis as the methodical approach. The objects in this case are the individual questions with the competency domain as the grouping variable. The number of points scored by the test-takers for the questions (response patterns) are the predictors. One problem here is that more people have participated than there are questions. Analogous to the approach taken for a principal component regression, data reduction was therefore carried out by determining the principal components. Linear discriminant analyses were then carried out with a reduced number of principal components. This "discriminant analysis of the principal components" (DAPC) is used for similar reasons in genetic analyses in which the number of predictors exceeds the number of objects to be classified ([8], see also note 2 below).

The first analytical step (determining the principal components) does not involve any distributional assumptions. The extraction of the principal components serves to reduce the original data to a few components in the response patterns which best approximate the data according to the least squares criterion. Fisher's linear discrim-

inant analysis is a special case of linear discriminant analysis in which the a priori group sizes are assumed to be equal. In this special case, no normal distributional assumption is made for the analysis. Calculation of the p-values for the group comparisons (see below) is done using t-tests that in principle are based on normal distributional assumptions but are known to be robust to violations of these assumptions (we also performed a non-parametric test by means of a randomizing test whose results are substantially identical to the t-test presented here and are therefore not reported here for reasons of clarity).

One question associated with the method of factor analysis and one that generally cannot be answered satisfactorily is the question concerning the determination of the number of principal components to be used. For the different SKPT, no consistent value over the years has been yielded by scree plots [21] or from Onatski's method [19], which is why we extracted ten principal components, meaning twice as many components as competency domains (the following results are less sensitive in relation to the number of the extracted components; analyses with more than six components yield nearly identical results).

To answer the question if the individual competency domains are different from the entirety of all other domains, a one-against-the-rest analysis and a pairwise classification of all competency domains against each other (one-against-one) were carried out [10], [27]. Tests of signifi-

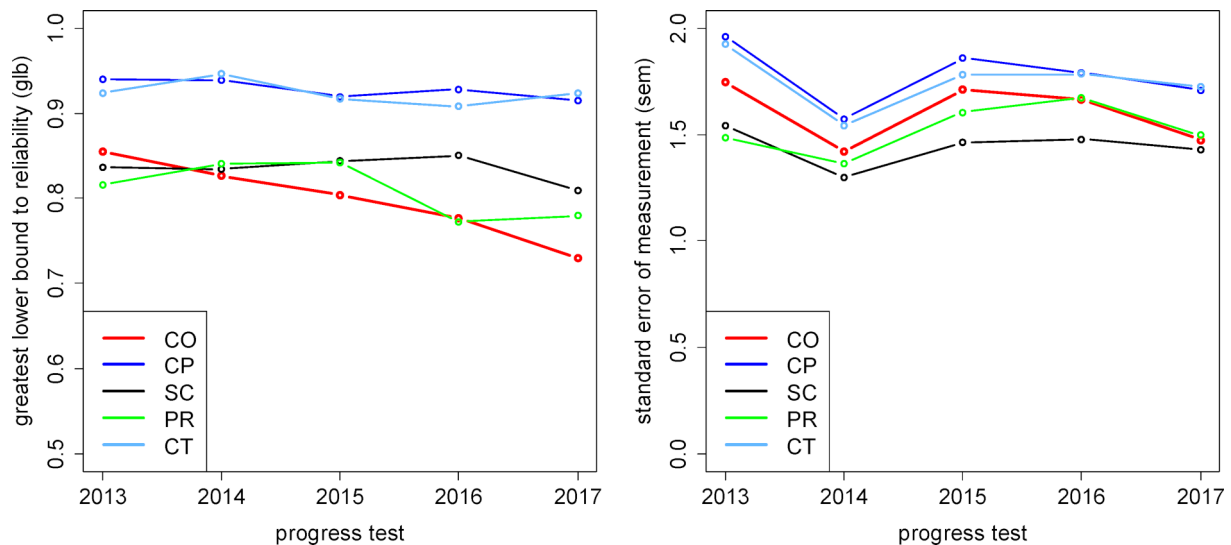


Figure 1: Reliability glb (left) and standard error of measurement sem (right) for the competency domains in the progress tests 2013–2017.

cance were done in each case by combining the individual p-values from the five progress tests for the years 2013 to 2017 using Fisher's method (Fisher's combined probability test). Significance was defined as $\alpha = 0.05$.

3. Results

3.1. Reliability and measurement precision

The number of questions, reliability coefficients glb, and the corresponding standard error of measurements (sem) are listed in table 5; a visual depiction of glb and sem over the years is presented in figure 1.

Practical clinical competence (CP) and theoretical clinical competence (CT) are measured as nearly stable with a high reliability (more than 0.90); (one notes that the number of the questions in these two competency domains is higher than in the other competency domains, see the blueprint). Scientific competence (SC) is likewise stable over time with a reliability of around 0.80–0.85. For professional decision-making competence (PR) and communicative competence (CO), a decrease to a reliability below 0.80 is seen for the last two years, most distinctly for communicative competence (CO) with a reliability of 0.73 for the 2017 progress test (see figure 1, left diagram).

For the standard error of measurement (see figure 1, right diagram), a distinct reduction in 2014 is visible that can be traced to the fewer number of questions in comparison to the SKPT for 2013. Afterward, the standard error of measurements for all competency domains remain virtually the same since 2015. Here, too, varying numbers of questions in the competency domains must be taken into account; longer scales also have a larger absolute sem. This is also reflected in figure 1 (left): practical clinical competence (CP) and theoretical clinical competence (CT) are the longest scales (with 30 ques-

tions each on the blueprint), and scientific competence (SC) has the shortest scale with only 18 questions.

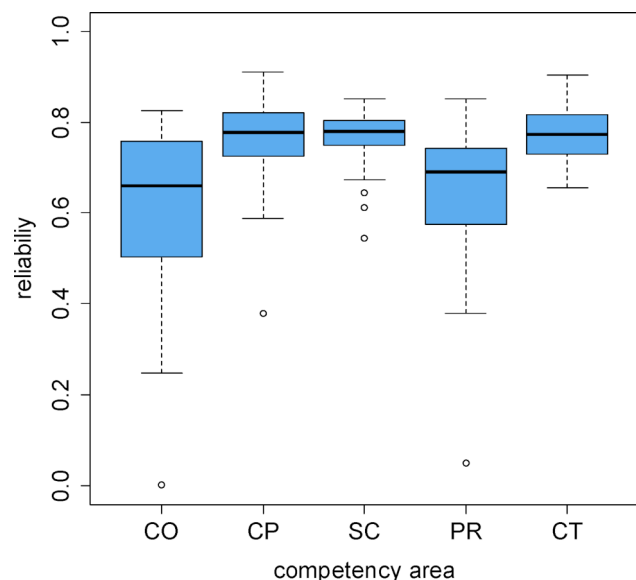


Figure 2: Reliabilities within year of study for the individual competency domains. Each box represents the 30 individual reliabilities of the five progress tests and six levels of study reflected in years of study.

The median values for the reliabilities in the subpopulations at the same semester level are presented in table 6 according to both year of study and competency domain. In addition, figure 2 shows the distribution of these 30 individual reliabilities for the different competency domains. While the measurement reliabilities for the majority of reliabilities for the practical clinical, theoretical clinical, and scientific competency domains (CP, CT, SC) are above 0.7 (CP 83%, SC 80%, CT 90%), these percentages are distinctly lower for communicative competence (CO) with 37% and for professional decision-making (PR) with 47%. In particular, for these individual SKPTs and semester levels it must be established that there are low

Table 5: Number of questions (n_A), reliability (coefficient glb) and standard error of measurement (sem) for the competency domains in the SKPTs for 201 –2017. Due to the exclusion of individual questions in the post-review, the numbers of questions per domain are sometimes less than intended in the blueprint (see table 2).

Comp. domain	Progress test														
	2013			2014			2015			2016			2017		
	n_A	glb	sem	n_A	glb	sem	n_A	glb	sem	n_A	glb	sem	n_A	glb	sem
CO	26	0.85	1.75	21	0.83	1.42	22	0.80	1.71	21	0.78	1.66	21	0.73	1.47
CP	38	0.94	1.96	28	0.94	1.57	30	0.92	1.86	28	0.93	1.79	28	0.92	1.71
SC	18	0.84	1.54	17	0.83	1.30	17	0.84	1.46	18	0.85	1.48	17	0.81	1.43
PR	20	0.82	1.48	19	0.84	1.36	19	0.84	1.60	20	0.77	1.67	19	0.78	1.50
CT	38	0.92	1.92	29	0.95	1.54	29	0.92	1.78	29	0.91	1.78	27	0.92	1.73

Table 6: Median values of the reliabilities for each year of study and competency domain from each of the five SKPTs for 2013-2017.

Year of study	Competency domain				
	CO	CP	SC	PR	CT
1	0.779	0.852	0.783	0.799	0.816
2	0.754	0.782	0.833	0.748	0.751
3	0.658	0.789	0.782	0.721	0.791
4	0.568	0.732	0.767	0.657	0.790
5	0.635	0.751	0.799	0.539	0.772
6	0.436	0.752	0.748	0.452	0.736

reliabilities under 0.4 (see figure 2). Low measurement reliabilities appear to be especially frequent for sixth-year students (percentages of the reliabilities over 0.7 for the first through fifth years of study: CO 44%, CP 84%, SC 84%, PR 56%, CT 96% (see table 6).

3.2. Delineation of the competency domains (“discriminant validity”)

3.2.1. Delineation of each competency domain in relation to all other domains

To provide examples, the results of the two-class discriminant analyses of a competency domain in relation to every other competency domain are presented in figure 3 and figure 4 for the progress tests given in 2013 and 2017 (the results for 2014 to 2016 are very similar). On the ordinate axes are the values of the discriminant function of the individual questions in the form of a box plot. Clear overlapping in the box plots of two competency domains points to construct overlapping in the associated questions. When the boxes are clearly separated this indicates that the two groups of questions, and hence the two competency domains, are distinct from each other. Thus, the questions belonging to scientific competence (SC) in all years form a clearly delineated cluster: it is highly probable that whoever answers one question from this competency domain correctly, also answers the other questions in this domain correctly (see figure 3, diagram

c and figure 4, diagram c). In contrast, there is a high degree of overlapping between practical clinical competence (CP) and theoretical clinical competence (CT) (see figure 3, diagrams b, e; and figure 4, diagrams b, e).

For inferential statistical testing to see if the individual competency domains differ from the entirety of all other domains, the individual p-values for 2013 through 2017 were combined in Fisher’s probability test (one notes that the individual p-values are not used for significance testing). There are significant values for all five domains (see table 7).

3.2.2. Pairwise delineation of the competency domains

To provide examples of the ten possible paired comparisons, we present the values of the discriminant function of the questions for the comparison between practical clinical competence (CP) and theoretical clinical competence (CT) in figure 5, and the values comparing scientific competence (SC) and theoretical clinical competence (CT) in figure 6.

We present the p-values, combined from the individual values for the different SKPTs, for all paired comparisons in table 8. All of these pairwise comparisons are significant, with the exception of the comparison between the practical clinical (CP) and the theoretical clinical (CT) competency domains.

These significant differences remain even after applying the Bonferroni-Holm method to adjust the ten tests: for

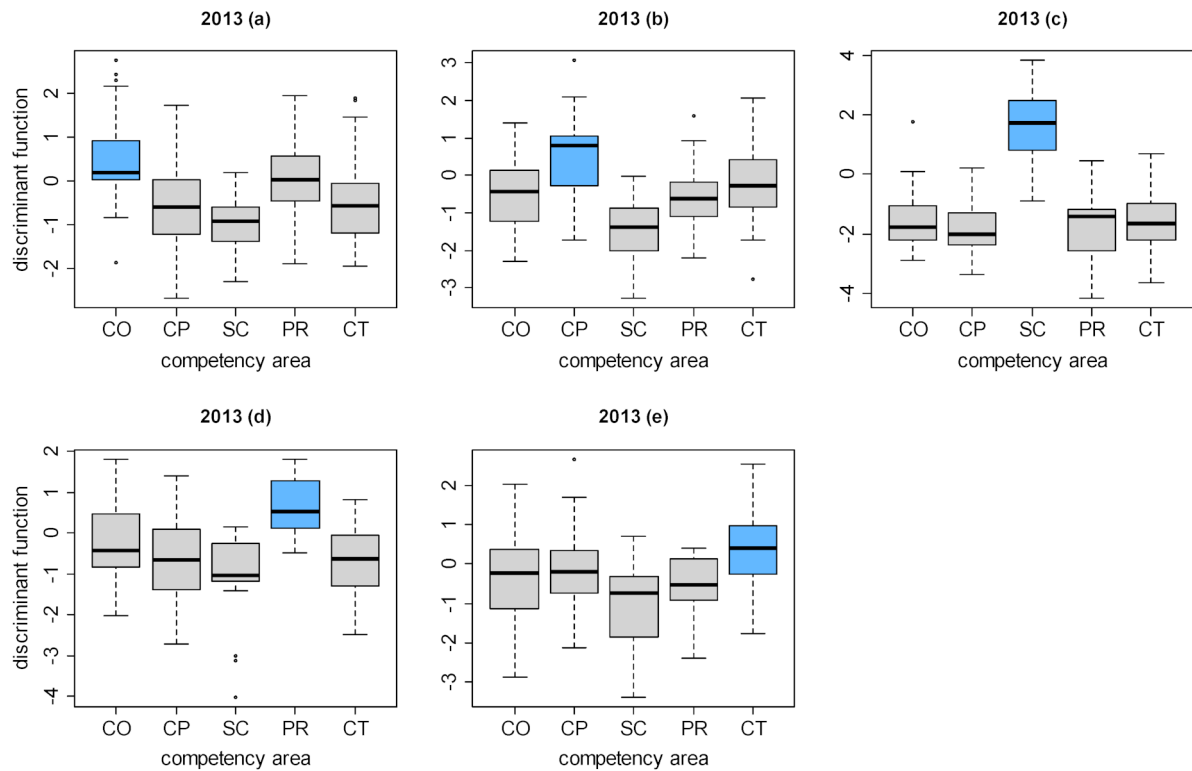


Figure 3: Distributions of the discriminant function values for the questions in the competency domains after performing two-class discriminant analysis for each domain in relation to all other domains (one-against-the-rest) for the 2013 progress test.

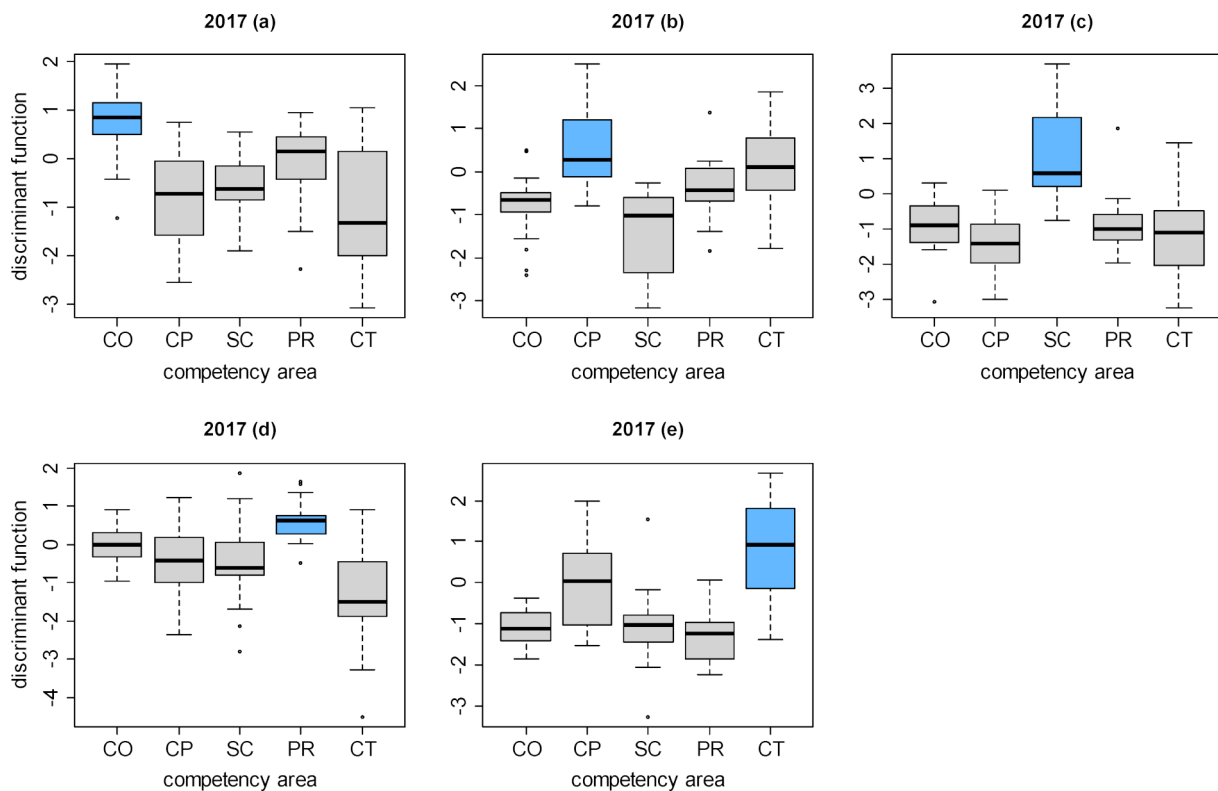


Figure 4: Distributions of the discriminant function values for the questions in the competency domains after performing two-class discriminant analysis for each domain in relation to all other domains (one-against-the-rest) for the 2017 progress test.

Table 7: p-values for the tests for difference between the values of the discriminant function of the questions in a competency domain of the progress tests 2013-2017 in relation to the questions for all other domains and for Fisher's combined probability test (FCPT).

Competency Domain	Progress test					FCPT
	2013	2014	2015	2016	2017	p
CO	0.055	0.017	<0.001	<0.001	0.002	<0.001
CP	<0.001	0.004	0.012	<0.001	0.026	<0.001
SCI	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
PR	0.010	0.002	0.002	0.001	0.032	<0.001
CT	0.030	0.076	0.018	<0.001	<0.001	<0.001

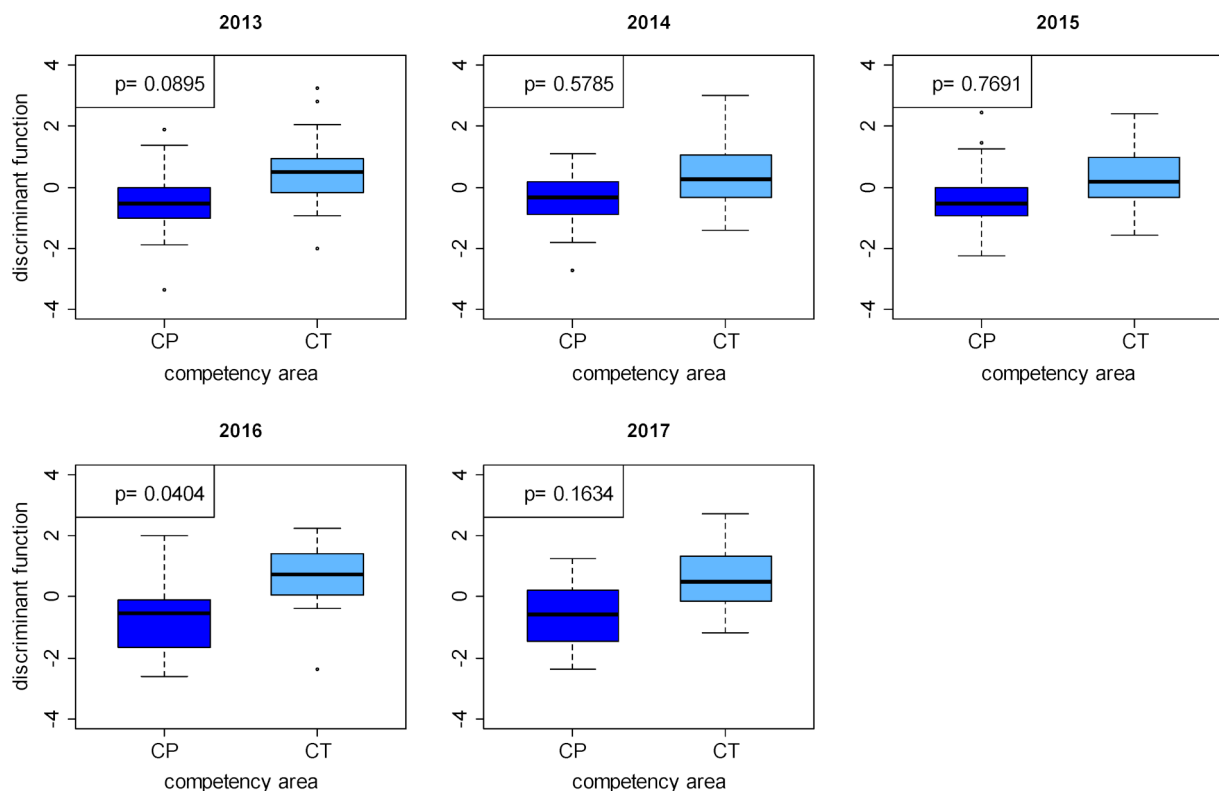


Figure 5: Distributions of the discriminant function values for the questions in the practical clinical competency domain and the theoretical clinical competency domain after performing two-class discriminant analysis for the progress tests 2013–2017.

this, the individual p-values of the tests are ordered in ascending order according to size ($p_{[1]} \leq p_{[2]} \leq \dots \leq p_{[10]}$). Significant are precisely those $p_{[k]}$ for which the inequalities $p_{[i]} \leq \alpha / (11 - i)$ are fulfilled for $\forall i \leq k$. If an inequality is not achieved for any smaller $p_{[i]}$, then no larger $p_{[k]}$ can still be considered significant.

For eight of the tests the result is $p < 0.001$, making these significant based on $p_{[1]} \leq 0.05/10$, $p_{[2]} \leq 0.05/9$, ..., $p_{[8]} \leq 0.05/3$; in addition, $p_{[9]} = 0.002 \leq 0.05/2 = 0.025$ is also significant. Only the p-value for the comparison between the competency domains CP and CT, which is $p_{[10]} = 0.091$, does not fulfill the condition $p_{[10]} \leq 0.05/1$.

4. Summary and discussion

The competency domains of practical clinical competence (CP), theoretical clinical competence (CT) and scientific competence (SC) have been reliably measured by five SKPTs over the years (reliability over 0.80). Not quite so satisfactory are the questions covering the domains of communicative competence (CO) and professional decision-making competence (PR), with reliabilities still over 0.73. The questions for the different competency domains also represent empirically different domains. One exception is seen in the groups of questions covering practical clinical and theoretical clinical competence, between which no delineation can be empirically demonstrated ($p = 0.091$).

With few exceptions, the measurement reliabilities within the student cohorts by year of study show satisfactory

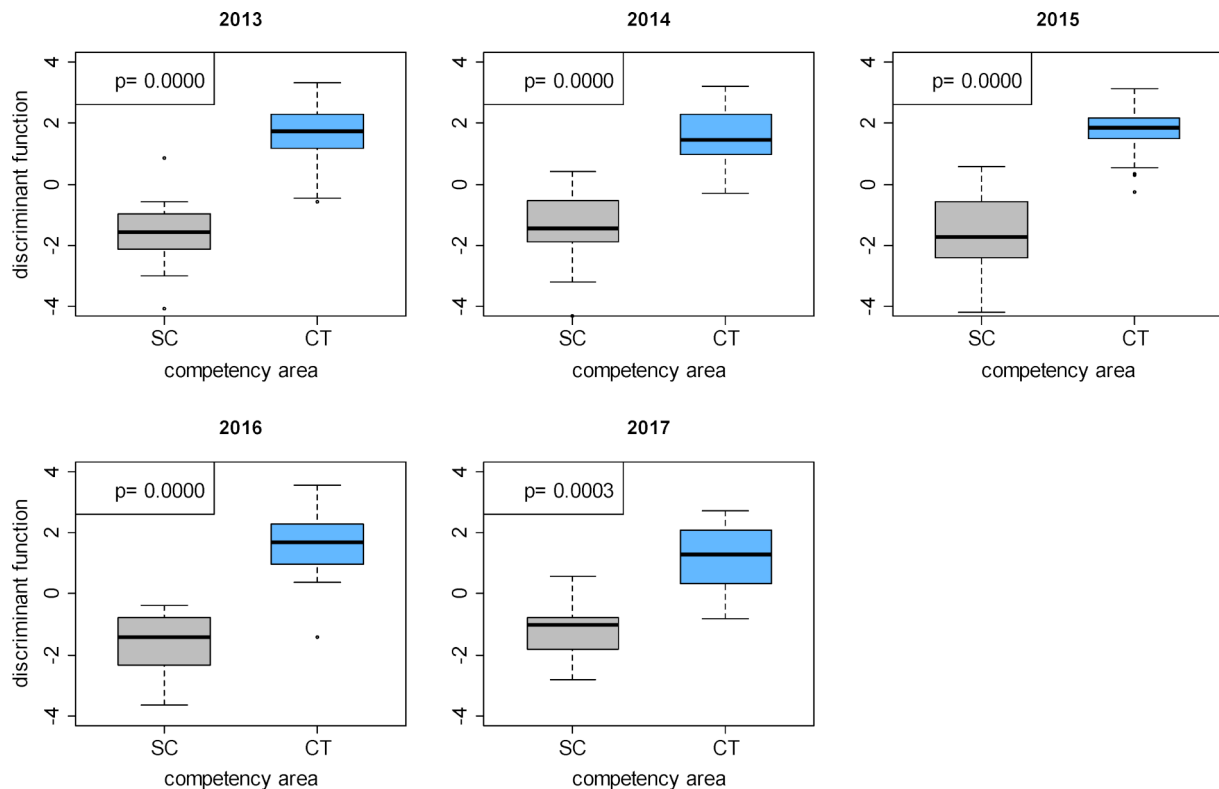


Figure 6: Distributions of the discriminant function values for the questions in the scientific competence domain and the theoretical clinical competence domain after performing two-class discriminant analysis for the progress tests 2013–2017.

Table 8: Fisher's combined probability test (FCPT) of the pairwise tests for difference in the values of the discriminant function between the questions of two competency domains.

	Competency domain				
	CO	CP	SC	PR	CT
CO	–				
CP	<0.001	–			
SC	<0.001	<0.001	–		
PR	0.002	<0.001	<0.001	–	
CT	<0.001	0.091	<0.001	<0.001	–

values for the practical clinical, theoretical clinical and scientific competency domains (CP, CT, SC). Limitations are seen in the reliabilities of the domains communication and professional decision making (CO, PR), which frequently still do not reach 0.7. However, it must be noted that these domains are covered by only 22 and 20 questions, respectively, numbers that are generally inadequate for achieving a sufficiently high reliability using Type A MCQs even on the well-prepared summative, subject-specific tests given at medical schools.

It is thus shown that on competency-based knowledge tests the generation of questions to measure practically relevant knowledge is also possible with a limited number of questions and – using the terminology of Cronbach and Meehl [2] – represent the different constructs in the blueprint intended for the competency domains.

The exceptions to this are the practical clinical and theoretical clinical competency domains (CP/CT), which, al-

though they differ significantly from the other competency domains, do not differ from each other. Because the discriminant analysis applied here explicitly allows construct overlap, the fact that no delineation can be drawn between these two domains cannot serve as an explanation. The reason for this non-separability could be rules that are too vague to clearly assign a question to only one of the two domains. An alternative explanation could be that within the context of medical expertise these two domains are very strongly connected in terms of content and that this knowledge is largely acquired by students at parallel points in time. Despite the semantic differences of each domain, this would lead to no detection of an empirical difference in the test question responses: whoever has the knowledge to answer the questions in one domain also has the knowledge to answer the questions in the other.

Discriminant analysis of the principal components (DAPC) proved itself to be methodically suitable for empirically tracing the underlying structure of the blueprint axis for the competency domains.

One consequence arising from these results for future progress tests should be an attempt to improve the measurement reliability for competence in communicative (CO) and professional decision-making (PR). This could be achieved by increasing the number of questions for each domain. To avoid enlarging the scope of the SKPT, it would be conceivable to reduce the number of questions asked in the domains of practical clinical competence (CP) and theoretical clinical competence (CT). Delineation of the content of these two domains should also be verified. If empirical discrimination remains impossible to determine, then these two domains could be combined in the feedback given to students.

Notes

1. Parts of the study concerning the 2013 and 2014 progress tests were reported at the 2014 GMA conference in Hamburg and at RIME 2015 in Munich [14], [15].
2. The authors are unaware of an original source on discriminant analysis of principal components; the earliest mention we found is in a paper on the distribution of larger mammals in a national park in Tanzania [16].

Funding

This study was undertaken within the scope of the MERLIN II project that is funded by the Federal Ministry of Education and Research (01PL17011C).

Competing interests

The authors declare that they have no competing interests.

Attachments

Available from

<https://www.egms.de/en/journals/zma/2020-37/zma001299.shtml>

1. Attachment_1.pdf (260 KB)
Examples of competency-specific questions from the 2017 competency-based progress test

References

1. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull.* 1959;56(2):81-105. DOI: 10.1037/h0046016
2. Cronbach LJ, Meehl P E. Construct validity in psychological tests. *Psychol Bull.* 1955;52(4):281-302. DOI: 10.1037/h0040957
3. Frank JR, Snell LS, Cate OT, Holmboe ES, Carraccio C, Swing SR, Harris P, Glasgow NJ, Campbell C, Dath D, Harden RM, Iobst W, Long DM, Mungroo R, Richardson DL, Sherbino J, Silver I, Taber S, Talbot M, Harris KA. Competency-based medical education: theory to practice. *Med Teach.* 2010;32(8):638-645. DOI: 10.3109/0142159X.2010.501190
4. Hahn EG, Fischer MR. Nationaler Kompetenzbasierter Lernzielkatalog Medizin (NKLM) für Deutschland: Zusammenarbeit der Gesellschaft für Medizinische Ausbildung (GMA) und des Medizinischen Fakultätentages (MFT). *GMS Z Med Ausbild.* 2009;26(3):Doc35. DOI: 10.3205/zma000627
5. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency based medical education. *Med Teach.* 2010;32(8):676-682. DOI: 10.3109/0142159X.2010.500704
6. Hughes DJ. Psychometric validity: Establishing the accuracy and appropriateness of psychometric measures. In: Irwing P, Booth T, Hughes DJ, editors. *Wiley handbook of psychometric testing: A multidisciplinary approach to survey, scale, and test development.* Hoboken, NJ: Wiley; 2018. DOI: 10.1002/9781118489772.ch24
7. Jackson, P, Agunwamba, C. Lower bounds for the reliability of the total score on a test composed of non-homogeneous items I: Algebraic lower bounds. *Psychomet.* 1977;42(4):567-578. DOI: 10.1007/BF02295979
8. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Gen.* 2010;11:94. DOI: 10.1186/1471-2156-11-94
9. Lord FM, Novick MR. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley; 1968.
10. Lotte F, Bougrain L, Cichocki A, Clerc M, Congedo M, Rakotomamonjy A, Yger F. A Review of Classification Algorithms for EEG-based Brain-Computer Interfaces: A 10-year Update. *J Neural Eng.* 2018;15(3):031005. DOI: 10.1088/1741-2552/aab2f2
11. Lurie, SJ. History and practice of competency-based assessment. *Med Educ.* 2012;46(1):49-57. DOI: 10.1111/j.1365-2923.2011.04142.x
12. McGill DA, van der Vleuten CP, Clarke MJ. A critical evaluation of the validity and the reliability of global competency constructs for supervisor assessment of junior medical trainees. *Adv Health Sci Educ Theory Pract.* 2013;18(4):701-725. DOI: 10.1007/s10459-012-9410-z
13. McHarg J, Bradley P, Chamberlain S, Ricketts C, Searle J, McLachlan JC. Assessment of progress tests. *Med Educ.* 2005;39(2):221-227. DOI: 10.1111/j.1365-2929.2004.02060.x
14. Möltner A, Wagener S, Jünger J. Empirische Struktur des kompetenzbasierten studentischen Progresstests: Reliabilität und diskriminante Validität von Kompetenzbereichen. In: *Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA).* Hamburg, 25.-27.09.2014. Düsseldorf: German Medical Science GMS Publishing House; 2014. DocV443. DOI: 10.3205/14gma308
15. Möltner A, Wagener S, Timbil S, Gornostayeva M, Jünger J. Empirical Structure of a Competency-Based Progress Test. In: *4th Research in Medical Education (RIME) Symposium 2015.* München, 19.-21.03.2015. Düsseldorf: German Medical Science GMS Publishing House; 2015. DocS1B3. DOI: 10.3205/15rime14
16. Morton-Griffiths M. The numbers and distribution of large mammals in Ruaha National Park, Tanzania. *E Afr Wildl J.* 1975;13:121-140. DOI: 10.1111/j.1365-2028.1975.tb00127.x

17. Newton PE, Shaw SD. Validity in educational and psychological assessment. Los Angeles: Sage; 2014. DOI: 10.4135/9781446288856
18. Nouns ZM, Georg W. Progress testing in German speaking countries. *Med Teach*. 2010;32(6):467-470. DOI: 10.3109/0142159X.2010.485656
19. Onatski A. Testing hypotheses about the number of factors in large factor models. *Econometr*. 2009;77(5):1447-1479. DOI: 10.3982/ECTA6964
20. Patterson F, Zibarras L, & Ashworth V. Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. *Med Teach*. 2016;38(1):3-17. DOI: 10.3109/0142159X.2015.1072619
21. Raïche G, Walls TA, Magis D, Riopel M, Blais JG. Non-graphical solutions for Cattell's scree test. *Methodol*. 2013;9(1):23-29. DOI: 10.1027/1614-2241/a000051
22. Ravesloot CJ, Van der Schaaf MF, Muijtens AMM, Haaring C, Kruitwagen CL, Beek FJ, Bakker J, Van Schaik JP, Ten Cate TJ. The don't know option in progress testing. *Adv Health Sci Educ*. 2015;20(5):1325-1338. DOI: 10.1007/s10459-015-9604-2
23. Schindler C, Bauer J, Strasser A, Schlomske-Bodenstein N, Seidel T, Prenzel M. Prüfungen als Indikator für den Studienerfolg. In: Berthold C, Jorzik B, Meyer-Guckel V, editors. *Handbuch Studienerfolg*. Essen: Edition Stifterverband; 2015. p.62-79.
24. Schuwirth LW, van der Vleuten CP. The use of progress testing. *Perspect Med Educ*. 2012;1(1):24-30. DOI: 10.1007/s40037-012-0007-2
25. Sijtsma K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*. 2009;74(1):107-120. DOI: 10.1007/s11336-008-9101-0
26. Steinhäuser J, Chenot JF, Roos M, Ledig T, Joos S. Competence-based curriculum development for general practice in Germany: a stepwise peer-based approach instead of reinventing the wheel. *BMC Res Notes*. 2013;6(1):314. DOI: 10.1186/1756-0500-6-314
27. Tax DM, Duin RP. Using two-class classifiers for multiclass classification. In: *Proceedings of the 16th International Conference on Pattern Recognitions*. Vol. 2; 2002 Aug 11-15; Quebec, Kanada. IEEE Press; 2002. p.124-127. DOI:10.1109/ICPR.2002.1048253
28. Van der Vleuten CP, Verwijnen GM, Wijnen WH. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Med Teach*. 1996;18(2):103-109. DOI: 10.3109/01421599609034142
29. Wagener S, Fleig A, Möltner A. Warum sind im studentischen kompetenzorientierten Progresstest in jedem Studienjahr etwa 25% der Antworten falsch? In: *Gemeinsame Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA) und des Arbeitskreises zur Weiterentwicklung der Lehre in der Zahnmedizin (AKWLZ)*. Münster, 20.-23.09.2017. Düsseldorf: German Medical Science GMS Publishing House; 2017. Doc157. DOI: 10.3205/17gma157
30. Wagener S, Gaitzsch E, Brass K, Heid J, Herrmann L, Günther J, Ney S, Müller A, Dikmen HO, Zimmermann K, Yilmaz OK, Wittstock F, Alhalabi O, Park J, Harapan BN, Kollmeier B, Ronellenfisch L, Mayer J, Baumann T, Daunert L, Brüstle P, Mohr D, Schüttelz-Brauns K, Van der Beken A, Jünger J, Möltner A. Videofragen im studentischen kompetenzorientierten Progresstest. In: *Gemeinsame Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA) und des Arbeitskreises zur Weiterentwicklung der Lehre in der Zahnmedizin (AKWLZ)*. Münster, 20.-23.09.2017. Düsseldorf: German Medical Science GMS Publishing House; 2017. Doc110. DOI: 10.3205/17gma110
31. Wagener S, Möltner A, Fleig A, Feistner L, Heid J, Brass K, Holz T, Weber M, Pflaum P, Rogg D, Kellermann F, Berg L, Breithaupt MH, Dehmel L, Grad A, Xiang Jin J, Hai-Ning Lu K, Müller A, Rinawi T, Shang V, Zimmermann K, Alhalabi O, Park J, Grupp M, Klauth A, Lepper A, Lichnock Z, Mayer J, Hollmann A, Meuth C, Siegel F, Peitz N, Brüstle P, Mohr D, Schüttelz-Brauns K, Würth G, Jünger J, Burkert M. "Was will ich prüfen?" - Erweiterung des Blueprints im studentischen kompetenzorientierten Progresstest. In: *Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA)*. Wien, 19.-22.09.2018. Düsseldorf: German Medical Science GMS Publishing House; 2018. Doc19.5. DOI: 10.3205/18gma338
32. Wagener S, Möltner A, Timbil S, Fleig A, Feistner L, Heid J, Brass K, Burkert M. "Da bin ich mir sicher" - Confidence rating im studentischen kompetenzorientierten Progresstest [Bericht über Forschungsergebnisse]. In: *Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA)*. Wien, 19.-22.09.2018. Düsseldorf: German Medical Science GMS Publishing House; 2018. Doc15.2. DOI: 10.3205/18gma067
33. Wagener S, Möltner A, Timbil S, Gornostayeva M, Schultz JH, Brüstle P, Mohr D, Van der Beken A, Better J, Fries M, Gottschalk M, Günther J, Herrmann L, Kreisel C, Moczek T, Illg C, Jassowicz A, Müller A, Niesert M, Strübing F, Jünger J. Development of a competencybased formative progress test with student-generated MCQs: Results from a multi-centre pilot study. *GMS Z Med Ausbild*. 2015;32(4):Doc46. DOI: 10.3205/zma000988
34. Wrigley W, van der Vleuten CP, Freeman A, Muijtens A. A systemic framework for the progress test: strengths, constraints and issues: AMEE Guide No. 71. *Med Teach*. 2012;34(9):683-697. DOI: 10.3109/0142159X.2012.704437
35. Zmud RW, Sampson JP, Reardon RC, Lenz JG, Byrd TA. Confounding Effects of Construct Overlap: An Example from IS User Satisfaction Theory. *Inform Technol People*. 1994;7(2):29-45. DOI: 10.1108/09593849410074061

Corresponding author:

Dr. phil. Andreas Möltner
Medical Faculty of Heidelberg, Baden-Württemberg Center of Excellence for Assessment in Medicine, Im Neuenheimer Feld 346, D-69120 Heidelberg, Germany
andreas.moeltner@med.uni-heidelberg.de

Please cite as

Möltner A, Wagener S, Burkert M. Measuring competency-relevant knowledge in the competency-oriented student progress test. *GMS J Med Educ*. 2020;37(1):Doc6.
DOI: 10.3205/zma001299, URN: urn:nbn:de:0183-zma0012993

This article is freely available from

<https://www.egms.de/en/journals/zma/2020-37/zma001299.shtml>

Received: 2019-02-04

Revised: 2019-08-01

Accepted: 2019-10-14

Published: 2020-02-17

Copyright

©2020 Möltner et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 License. See license information at <http://creativecommons.org/licenses/by/4.0/>.

Erfassung kompetenzrelevanten Wissens im studentischen kompetenzorientierten Progresstest

Zusammenfassung

Hintergrund: Seit dem Jahr 2013 wird an einer Reihe medizinischer Fakultäten ein studentischer kompetenzorientierter Progresstest (SKPT) durchgeführt. Die Erstellung der Fragen erfolgt auf Basis eines zweidimensionalen Blueprints, dessen eine Achse aus den auf dem Kompetenzmodell des NKLM basierenden fünf Kompetenzbereichen „Kommunikative Kompetenz“ (KO), „Klinisch-praktische Kompetenz“ (KP), „Klinisch-theoretische Kompetenz“ (KT), „Wissenschaftskompetenz“ (WI) und „Professionelle Handlungskompetenz“ (PH) besteht. Die Rückmeldung an die teilnehmenden Studierenden erfolgt u. a. differenziert nach diesen Bereichen. Ziel der Studie ist, zu prüfen, ob

1. die nach Kompetenzbereichen differenzierten Ergebnisse eine hinreichende Messgenauigkeit aufweisen und
2. ob die Ergebnisse zu verschiedenen Bereichen auch unterschiedliche Informationen enthalten.

Methoden: Untersucht wurden die SKPTs der Jahre 2013 bis 2017 mit zusammen 3027 Teilnehmern. Die Messgenauigkeit wurde mit dem Koeffizient g_{lb} und dem Standardmessfehler bestimmt, zum Nachweis der Differenziertheit der Kompetenzbereiche wurde eine Diskriminanzanalyse der Hauptkomponenten eingesetzt.

Ergebnisse: Die Reliabilität der Kompetenzbereiche war in allen Progresstests über 0,8, Ausnahmen hiervon waren in zwei der Tests KO und PH mit einer Reliabilität von 0,7–0,8. Die Ergebnisse aller Bereiche unterschieden sich hinsichtlich ihres Informationsgehalts jeweils vom Gesamt der anderen Bereiche, gleiches gilt mit Ausnahme von KP und KT für sämtliche paarweisen Vergleiche.

Diskussion: Die nach Kompetenzbereichen differenzierte Rückmeldung der Leistungen im SKPT erfüllt im Wesentlichen die Anforderungen an Messzuverlässigkeit und Eigenständigkeit. Eine Verbesserung der Messgüte bei den Bereichen KO und PH sowie eine bessere Differenzierung der Bereiche KP und KT ist wünschenswert.

Schlüsselwörter: Progresstest, Reliabilität, Diskriminanzvalidität

Andreas Möltner¹
Stefan Wagener¹
Mirka Burkert¹

1 Medizinische Fakultät
Heidelberg,
Kompetenzzentrum für
Prüfungen in der Medizin
Baden-Württemberg,
Heidelberg, Deutschland

1. Hintergrund

Eine kompetenzbasierte medizinische Ausbildung („competency based medical education“ – CBME) ist insbesondere in den letzten 20 Jahren in den Fokus des Interesses der Curriculumsentwicklung an den medizinischen Fakultäten und der Gesundheitspolitik gerückt [3], [26]. Konsequenterweise muss sich dies auch in den Leistungserfassungen widerspiegeln [5]. Dies hat zu einer wahren Flut an Veröffentlichungen zum Thema kompetenzbasierten Prüfens geführt (man vergleiche hierzu die kritische Diskussion in [11]), wobei insbesondere praktische und arbeitsplatzbasierte Verfahren zur Erfassung von Kommunikationskompetenz, Handlungskompetenz u. ä. im Zentrum stehen. Weniger Aufmerksamkeit haben

Prüfungen mittels traditioneller Multiple-Choice-Aufgaben in diesem Zusammenhang erfahren, da – ausgehend von Kompetenzen als primär handlungsbezogen – der reinen Erfassung von Wissen eine geringere Bedeutung zugesprochen wird. Dennoch ist Wissen eine wesentliche Voraussetzung kompetenter Handlungen, welches mittels „kompetenzorientierter Wissenstests“ [23] erfasst werden kann. Hierfür bieten MC-Fragen nach wie vor erhebliche Vorteile: Auf Grund der im Allgemeinen kurzen Bearbeitungszeit einer Aufgabe kann in einer Prüfung eine hohe Zahl von Fragen gestellt werden, die damit – relativ zum Zeitaufwand – eine deutlich bessere Abdeckung und Repräsentativität für den zu prüfenden Lehrinhalt erlaubt als aufwändig zu korrigierende offene schriftliche Frageformate oder praktische Prüfungsformen. Zudem sind Objektivität und eine hohe Reliabilität leichter zu erreichen als etwa mit arbeitsplatzbasierten Prüfungen, bei

denen die Gefahr besteht, dass ihre augenscheinlich höhere Validität durch ungenügende Standardisierung zunichte gemacht wird.

Im Rahmen des vom BMBF geförderten Verbund-Projekts „Medical Education Research – Lehrforschung im Netz BW“ (MERLIN, <http://www.merlin-bw.de/>) wurde im Jahr 2013 vom Kompetenzzentrum für Prüfungen in der Medizin/Baden-Württemberg ein „studentischer kompetenzorientierter Progresstest“ (SKPT) konzipiert.

Progresstests sind in der medizinischen Ausbildung anerkannte und eingesetzte Verfahren, um den Lernfortschritt im Verlauf des Studiums abzubilden [28]. Dazu wird Studierenden aller Ausbildungsstufen (Studienjahre) derselbe Test vorgelegt. Bei hinreichender Sicherung der Äquivalenz der Tests (gleiche Schwierigkeit) in aufeinanderfolgenden Jahren, kann für den Studierenden sein Lernfortschritt im Laufe seiner Ausbildung abgebildet und entsprechendes Feedback gegeben werden. Progress-tests werden auch in Deutschland an verschiedenen Fakultäten durchgeführt [18] und haben Progresstests haben vornehmlich zwei Funktionen: Einerseits geben sie Studierenden ein fortlaufendes Feedback über ihren individuellen Wissensstand im Studium, andererseits bieten sie für die Fakultäten die Möglichkeit, Curricula zu monitoren, Leistungsfortschritte in unterschiedlichen Kohorten zu beobachten und verschiedenen gestalteten Curricula zu vergleichen [24], [34].

Die Aufgabenzusammenstellung des SKPT beruht auf einem zweidimensionalen Blueprint, dessen eine Achse aus den inhaltlich in acht Gruppen zusammengefassten Fächern der Approbationsordnung und dessen andere Achse aus den in fünf Kompetenzbereiche gruppierten Einzelkompetenzen des Nationalen kompetenzbasierten Lernzielkatalogs Medizin (NKLM) [4] besteht. Der Blueprint wurde von drei interfakultären und interdisziplinären Expertengruppen gebildet, die die Aufgabe hatten, die Gebiete des NKLM in Clustern von „Kompetenzbereichen“ zusammenzustellen sowie für die Leistungsnachweise der ÄAppO „Fächergruppen“ zu bilden, wobei diese Gruppen jeweils vorklinische und klinische Fächer möglichst ausgewogen enthalten sollten (siehe Tabelle 1 und Tabelle 2).

Eine Besonderheit dieses Progresstests ist, dass die Aufgaben von Studierenden erstellt werden, die hierfür in mehrfach jährlich stattfindenden Workshops geschult werden. Die teilnehmenden Studierenden erstellen dabei anhand der im NKLM definierten Kompetenzen Fragen mit Bezug zu den Fächergruppen, um die Zellen des Blueprints in Tabelle 2 zu füllen. Die Zahl der vorhandenen Fragen je Zelle wird fortlaufend dokumentiert, so dass insbesondere gegen Ende des Prozesses der Aufgabenerstellung gezielt nur noch Fragen für nicht gefüllte Zellen formuliert werden müssen. Für eine detaillierte Darstellung sei auf [33] verwiesen.

Der Progresstest wird (seit 2015 in Kooperation mit dem Institut für Kommunikations- und Prüfungsforschung gGmbH) jährlich im November/Dezember durchgeführt. Teilnehmen können alle Studierenden der Humanmedizin aus Medizinischen Fakultäten, an denen der SKPT ange-

boten wird, im Jahr 2017 waren dies 16 Medizinische Fakultäten: Dresden, Erlangen-Nürnberg, Freiburg, Gießen, Hannover, Heidelberg, Homburg, Krems (A), Leipzig, Magdeburg, Mannheim, Marburg, LMU München, Tübingen, Ulm und Witten/Herdecke. Mit Ausnahme der Karl-Landsteiner-Privatuniversität in Krems war die Teilnahme freiwillig.

Der Test besteht aus 120 MC-Fragen des Typs A (eine korrekte Antwort aus vier oder fünf Antwortoptionen) mit der zusätzlichen Antwortoption „weiß nicht“ sowie 10 Situational-Judgement-Aufgaben (SJT) zur Erfassung sozialer Kompetenzen [20]. Ausnahme hiervon ist der erste Progresstest 2013 mit 144 Typ A-Aufgaben, jedoch ohne SJT.

Die „weiß nicht“-Option wird in Progresstests wie auch in anderen formativen Tests häufig eingesetzt, um es den Teilnehmern zu ermöglichen, explizit ihr Wissensdefizit zu dokumentieren und das Raten unter den verfügbaren Antwortoptionen vermeiden zu können [13], [22].

Beispiel einer Frage aus dem Kompetenzbereich „Klinisch-theoretische Kompetenz“ (KT) im SKPT 2017:

Sie behandeln auf Ihrer pädiatrischen Station ein zwölf-jähriges Kind mit Lungenentzündung, das schon häufig wegen Atemwegsbeschwerden und -infektionen in Behandlung war. Bei einer Ultraschalluntersuchung stellen Sie zudem einen Situs inversus, also eine gespiegelte Organausrichtung, fest. Daher senden Sie im Rahmen einer Bronchoskopie eine Biopsie in die Histologie. Welche Diagnose erwarten Sie sich von der histologischen Untersuchung?

1. Defekt der Zell-Zell-Kontakte
2. Defekt der Kinozilien
3. Ausbildung eines falschen Epithels
4. Fehlen von Oberflächendifferenzierungen
5. Defekt der Basalmembran
6. Weiß nicht

Die korrekte Antwort lautet hier B. Weitere Beispiele sind als Anhang 1 angefügt, sämtliche Fragen der SKPTs mit Erläuterungen finden sich auf der öffentlich zugänglichen Internetseite

<https://www.komp-pt.de/fragen-aus-dem-progresstest/>. Nach der Durchführung erhalten alle teilnehmenden Studierenden eine nach Fächergruppen und Kompetenzbereichen differenzierte Rückmeldung über ihre erbrachten Leistungen (erreichte Punktzahlen). Die Rückmeldung erfolgt absolut (kriteriumsbezogen), relativ im Vergleich zu den teilnehmenden Studierenden desselben Jahrgangs (normbezogen) sowie – bei mehrfacher Teilnahme – longitudinal zur Darstellung des Leistungszuwachses (Progress) im Vergleich zu den früheren Ergebnissen (ipsativ).

Tabelle 1: Blueprintachsen des SKPT: Zuordnung der Einzelkompetenzen des „Nationalen Kompetenzbasierten Lernzielkatalogs Medizin“ (NKLK) zu den Kompetenzbereichen des studentischen kompetenzbasierten Progresstests und Zuordnung der Fächer der ÄAppO zu den Fächergruppen.

Kompetenzen im NKLK	Kompetenzbereich im SKPT
Kollaborator, Kommunikator, Gesundheitsberater, Gesundheitsprävention	Kommunikative Kompetenz (KO)
Anwendung diagnostischer Verfahren, Klinisch-praktische Fertigkeiten, Notfallmaßnahmen, Therapeutisches Vorgehen	Klinisch-praktische Kompetenz (KP)
Gelehrte	Wissenschaftskompetenz (WI)
Professionell Handelnde, Ethik und Recht, Manager	Professionelle ärztliche Handlungskompetenz (PH)
Prinzipien normaler Funktion und Struktur, Pathophysiologische Mechanismen, Therapeutische Prinzipien, Diagnostischer Verfahren, Präventive Maßnahmen	Klinisch-theoretische Kompetenz (KT)

Fächer nach ÄAppO	Fächergruppe
Anatomie, Biologie, Rechtsmedizin, (Bildgebende Verfahren), Strahlenbehandlung und Strahlenschutz, Physik, Pathologie, klinisch pathologische Konferenz	I
Innere Medizin, Hygiene, Mikrobiologie und Virologie, Infektiologie und Immunologie, Physiologie	II
Chirurgie und Orthopädie, Urologie, HNO, Heilkunde, Augenheilkunde, Anästhesiologie, Notfallmedizin, Dermatologie und Venerologie	III
med. Soziologie, klin. Umweltmedizin, Arbeitsmedizin, Sozialmedizin, Epidemiologie, med. Biometrie und med. Informatik, Gesundheitsökonomie und -system sowie öffentl. Gesundheitswesen, Geschichte, Theorie und Ethik der Medizin, medizinische Terminologie	IV
med. Psychologie, Psychosomatische Medizin und Psychotherapie, Psychiatrie und Psychotherapie, Neurologie	V
Allgemeinmedizin, Medizin des Alterns und des alten Menschen, Schmerzmedizin, Palliativmedizin, Prävention, Rehabilitation, Physikalische Medizin und Naturheilverfahren	VI
Frauenheilkunde und Geburtshilfe, Kinderheilkunde, Humangenetik	VII
Biochemie/Molekularbiologie, Chemie, klinische Chemie und Laboratoriumsdiagnostik, klinische Pharmakologie und Pharmakotherapie, Pharmakologie und Toxikologie	VIII

Tabelle 2: Blueprint des studentischen kompetenzbasierten Progresstests 2014-2017 (zur Erläuterung der Kompetenzbereiche und Fächergruppen siehe Tab. 1).

Kompetenz-	Fächergruppen								Gesamt	Anteil (ca.)
bereich	I	II	III	IV	V	VI	VII	VIII		
KO	2	5	5	2	2	2	2	2	22	20.0%
KP	3	6	6	3	3	3	3	3	30	25.0%
WI	2	3	3	2	2	2	2	2	18	12.5%
PH	2	4	4	2	2	2	2	2	20	17.5%
KT	3	6	6	4	3	3	3	3	30	25.0%
Gesamt	12	24	24	12	12	12	12	12	120	
Anteil	10%	20%	20%	10%	10%	10%	10%	10%		

Ziel der Studie: Es soll geprüft werden, ob

1. die nach Kompetenzbereichen differenzierten Leistungserfassungen jeweils eine hinreichende Messgenauigkeit aufweisen sowie
2. durch die unterschiedlichen Aufgaben der Kompetenzbereiche empirisch voneinander unterschiedene Aufgabengruppen gebildet werden.

Beides ist insbesondere für die Nützlichkeit der Rückmeldungen an die teilnehmenden Studierenden von Bedeutung:

1. Rückmeldungen sind nur dann brauchbar, wenn die zu beurteilenden Leistungen zuverlässig erfasst sind.
2. Differenzierte Rückmeldungen machen nur dann Sinn, wenn die verschiedenen Einzelresultate auch Unterschiedliches beinhalten, also nicht redundant sind.

In der Terminologie von Campbell und Fiske [1] handelt es sich bei der Fragestellung (2) um den Nachweis der „Diskriminanzvalidität“ (auch „diskriminative Validität“). Der Begriff der „Validität“ war in den letzten 30 Jahren Gegenstand intensiver Diskussionen. Von vielen Autoren wird die Verwendung der unterschiedlichen Validitätsbegriffe (z. B. prädiktive, konvergente, diskriminative ... Validität) kritisch gesehen (eine eingehende Darstellung hierzu findet sich in [17]). In der vorliegenden Studie soll er dennoch in seiner „klassischen“ Bedeutung Verwendung finden: Die den Kompetenzbereichen zugeordneten Aufgaben bilden „Skalen“, welche unterschiedliche Leistungen messen (vgl. die Diskussion in [6]). Diese „Skalen“ sollten sich damit auch in den Antworten der Testteilnehmerinnen und -teilnehmer widerspiegeln, im Wesentlichen sollten Aufgaben des gleichen Bereichs ähnlich gut (oder schlecht) beantwortet werden.

2. Methoden

2.1. Datenerhebung

Die Durchführung des SKPT in den Jahren 2013-2017 erfolgte einmal jährlich. In den beiden ersten Jahren wurde der Test an den jeweiligen Fakultäten in Papierform durchgeführt, seit 2015 steht er online zur Verfügung. Die Ankündigung und Anmeldung der Studierenden sowie die Durchführung des Progresstests vor Ort erfolgte selbstständig seitens der beteiligten medizinischen Fakultäten. Die übergeordnete Koordination lag beim „Kompetenzzentrum für Prüfungen in der Medizin/Baden-Württemberg“ an der Medizinischen Fakultät Heidelberg. Genauere Details zur Durchführung finden sich in [33], [29], [30], [31], [32].

Der SKPT ist als formativer Test mit freiwilliger Teilnahme konzipiert. In der Fakultät in Krems ist der Test verpflichtend. Da sich in verschiedenen (hier nicht dargestellten) Analysen gezeigt hat, dass diese Studierendengruppe sich deutlich von den freiwilligen Teilnehmern aus den anderen Fakultäten unterscheidet, wurde diese Gruppe in den nachfolgenden Analysen nicht mit einbezogen.

Weiteres Einschlusskriterium war die Zahl der beantworteten Aufgaben. In der vorliegenden Studie wurden nur die Teilnehmer berücksichtigt, die mindestens 100 der 120 Aufgaben bearbeitet haben, d.h. die eine der vier oder fünf Antwortoptionen oder „weiß nicht“ angegeben haben (siehe Tabelle 3). Aus den beiden ersten SKPTs im Papierformat der Jahre 2013 und 2014 wurden nur vollständige Datensätze einbezogen. Eine Aufteilung der Teilnehmer nach Studienjahr findet sich in Tabelle 4. Die 10 Situational-Judgement-Aufgaben der Jahre 2014-2017 waren nicht Gegenstand der Studie.

Die Teilnehmer am SKPT willigten vor Durchführung des Tests ein, dass ihre Daten in pseudonymisierter Form zur Qualitätssicherung und für wissenschaftliche Zwecke verwendet werden dürfen.

2.2. Statistische Analysen

2.2.1. Bewertung der Aufgaben

Die nach dem Blueprint erstellten Aufgaben waren vom Typ A mit der zusätzliche Option „weiß nicht“. Bei Wahl der zutreffenden Antwort, wurde ein Punkt vergeben. Für die hier dargestellte Analyse wurden falsche Antworten und „weiß nicht“ sowie fehlende Antworten gleich behandelt und 0 Punkte vergeben (für alternative Bewertungsschemata, in denen „weiß nicht“ und Falschantworten unterschiedlich behandelt werden, s. [22]). Nach Durchführung des Tests erfolgte auf Grundlage von Teilnehmerkommentaren und der statistischen Auswertung eine zweite Begutachtung der Aufgaben (Post-Review). Stellte sich dabei heraus, dass bei einer Aufgabe mehrere Antworten zutreffend waren, erhielten die Teilnehmerinnen und Teilnehmer einen Punkt, wenn eine der korrekten Antworten gewählt wurde. Zwischen drei und acht Aufgaben in den SKPTs der Jahre 2013-2017 mussten nach dem Post-Review als fehlerhaft ausgeschlossen werden (siehe Tabelle 3).

2.2.2. Reliabilität und Messgenauigkeit

Zur Abschätzung der Reliabilität der Kompetenzbereiche wird die „greatest lower bound to reliability“ (glb) verwendet [7], [25]. Diese ist die algebraisch optimale Abschätzung der Reliabilität auch bei nicht-homogenen Skalen (in diesen Fällen liefert Cronbachs α als Maß der internen Konsistenz eine Unterschätzung der Reliabilität).

Die Reliabilität ist ein auf die Teilnehmerpopulation bezogenes relatives Maß der Messgenauigkeit, als absolutes Maß der Messgenauigkeit dient der Standardmessfehler, welcher sich aus der Reliabilität und der Standardabweichung der Skalenwerte der Teilnehmer errechnet [9].

Für den Vergleich mit Studierenden aus dem gleichen Studienjahr ist zusätzlich die Reliabilität bezogen auf diese Teilpopulationen von Relevanz. Hier sind gegenüber der Gesamtreliabilität niedrigere Werte zu erwarten, da (unter der Annahme annähernd gleicher Messfehler) die Varianz der Zahl der korrekten Antworten innerhalb der Teilnehmer des gleichen Studienjahrs niedriger als die

Tabelle 3: Format, Zahl der Items und Teilnehmerzahlen der SKPTs (ohne die Fakultät Krems) in der vorliegenden Untersuchung. In die Analyse wurden die Datensätze der Teilnehmer aufgenommen, die mindestens 100 Aufgaben beantwortet haben. In der letzten Spalte ist in Klammern die Zahl der Teilnehmer aufgeführt, die alle Aufgaben bearbeitet haben.

Progress-test	Format	Zahl der Fakultäten	MC-Items/ nach Post-Review ausgeschlossen	MC-Items in Analyse	Teilnehmer	Teilnehmer in Analyse (vollständig beantwortet)
2013	Papier	8	144 (-4)	140	434	433 (433)*
2014	Papier	8	120 (-6)	114	302	302 (302)
2015	Online	11	120 (-3)	117	1125	845 (832)
2016	Online	14	120 (-4)	116	1246	713 (694)
2017	Online	15	120 (-8)	112	1268	735 (726)

*Ein Teilnehmer mit fehlenden Angaben zum Studienjahr wurde ausgeschlossen

Tabelle 4: Aufteilung der Teilnehmer in der vorliegenden Untersuchung nach Studienjahren.

Progress-test	Studienjahr						Gesamt
	1	2	3	4	5	6	
2013	43	38	66	114	155	17	433
2014	82	27	54	43	81	15	302
2015	190	94	123	194	191	53	845
2016	118	84	113	177	140	81	713
2017	165	98	114	143	142	73	735
Summe	598	341	470	671	709	239	3028

aller Teilnehmer über die Studienjahre hinweg ist (vgl. hierzu [33]).

2.2.3. Abgrenzung der Kompetenzbereiche („Diskriminanzvalidität“)

Die Prüfung der Frage, ob die verschiedenen Kompetenzbereiche auch empirisch Unterschiedliches bezeichnen, gestaltet sich auf Grund des Konstruktionsprinzips des Progresstests etwas komplex: Die einzelnen Aufgaben sind nicht nur jeweils einem Kompetenzbereich sondern auch einem Fach oder einer Fachgruppe zugeordnet. Damit überlagern sich schon durch die Testkonstruktion bedingt die durch die beiden Achsen des Blueprint definierten Konstrukte („construct overlap“, nähere Erläuterungen hierzu finden sich z. B. in [35]). Daher ist nicht davon auszugehen, dass sich die Kompetenzbereiche mit faktoren- oder clusteranalytischen Verfahren direkt in den Daten abzeichnen (man vergleiche hierzu die faktoranalytische Untersuchung zur Validität und Reliabilität von Kompetenzkonstrukten in [12]).

Als methodischer Ansatz wurde deshalb das Verfahren der linearen Diskriminanzanalyse nach Fisher gewählt. Die Objekte sind in diesem Fall die einzelnen Aufgaben mit dem Kompetenzbereich als Gruppierungsvariable. Die erreichten Punktzahlen der teilnehmenden Personen bei den Aufgaben („Lösungsmuster“) sind die Prädiktoren. Dabei stellt sich das Problem, dass mehr Personen teilgenommen haben als Aufgaben vorhanden sind. Analog

zum Vorgehen bei einer Hauptkomponentenregression wurde deshalb eine Datenreduktion durch die Bestimmung von Hauptkomponenten vorgenommen. Die linearen Diskriminanzanalysen wurden danach mit einer verminderten Zahl von Hauptkomponenten durchgeführt. Diese „Diskriminanzanalyse der Hauptkomponenten“ („Discriminant analysis of principal components“, DAPC) wird aus ähnlichen Gründen z. B. im Bereich genetischer Analysen eingesetzt, bei der die Zahl der Prädiktoren die der zu klassifizierenden Objekte übersteigt ([8], vgl. auch Anmerkung 2).

Der erste Schritt der Auswertung (Bestimmung der Hauptkomponenten) beinhaltet keine Verteilungsannahmen. Die Extraktion der Hauptkomponenten dient zur Reduktion der Originaldaten auf wenige Komponenten in den Antwortmustern, die die Daten nach dem Kleinst-Quadrat-Kriterium bestmöglich approximieren. Die lineare Diskriminanzanalyse nach Fisher ist ein Spezialfall der linearen Diskriminanzanalyse, bei der die a priori-Gruppengrößen als gleich angenommen werden. In diesem Spezialfall wird keine Normalverteilungsannahme für die Daten getroffen. Die Bestimmung der p-Werte für die Gruppenvergleiche (s. u.) erfolgt mittels t-Tests, die zwar im Prinzip auf Normalverteilungsannahmen beruhen, jedoch bekanntlich robust gegenüber deren Verletzung sind (eine nicht-parametrische Testung mittels eines Randomisierungstests wurde ebenfalls durchgeführt, deren Ergebnisse sind substantiell mit den hier dargestell-

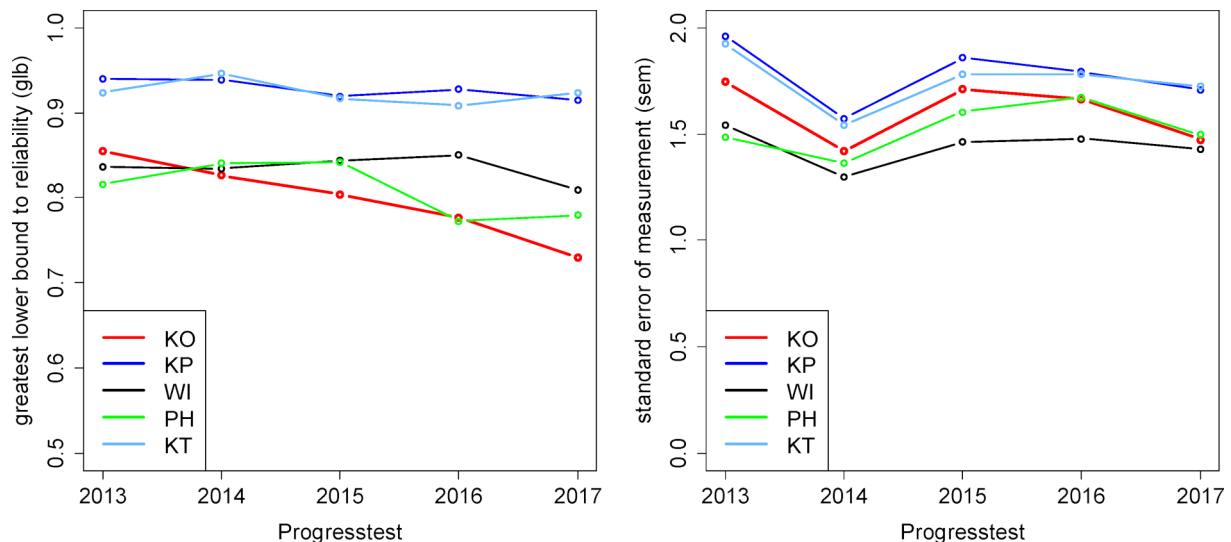


Abbildung 1: Reliabilität (links) und Standardmessfehler (rechts) der Kompetenzbereiche in den Progresstests 2013–2017.

ten t-Tests identisch und werden deshalb aus Gründen der Überschaubarkeit hier nicht berichtet).

Eine aus der Faktorenanalyse bekannte und allgemein nicht zufriedenstellend zu beantwortende Frage ist die nach der Bestimmung der Zahl der zu verwendenden Hauptkomponenten. Für die hier untersuchten Progress-tests ergab sich weder aus dem Scree-Test [21] oder dem Verfahren von Onatski [19] über die Jahre ein konsistenter Wert, weshalb zehn Hauptkomponenten, also doppelt so viele Komponenten wie Kompetenzbereiche, extrahiert wurden (die nachfolgend berichteten Ergebnisse sind wenig sensitiv gegenüber der Zahl der extrahierten Komponenten, es ergeben sich nahezu identische Resultate bei Analysen mit mehr als sechs Komponenten).

Zur Beantwortung der Frage, ob die einzelnen Kompetenzbereiche sich vom Gesamt aller anderen Bereiche abheben, wurde eine Analyse „one against the rest“ sowie eine paarweise Klassifikation aller Kompetenzbereiche gegeneinander („one against one“) durchgeführt [10], [27]. Signifikanztestungen erfolgten jeweils durch Kombination der einzelnen p-Werte aus den fünf Progresstests der Jahre 2013–2017 mittels der Methode von Fisher („Fisher's combined probability test“). Als Signifikanzniveau wurde $\alpha=0.05$ gewählt.

3. Ergebnisse

3.1. Reliabilität und Messgenauigkeit

In Tabelle 5 sind Aufgabenzahlen, Reliabilitätskoeffizienten glb und zugehörige Standardmessfehler sem aufgeführt, eine graphische Darstellung von glb und sem über die Jahre hinweg findet sich in Abbildung 1.

„Klinisch-praktische Kompetenz“ (KP) und „Klinisch-theoretische Kompetenz“ (KT) werden annähernd stabil mit einer hohen Reliabilität (über 0.90) erfasst (man beachte dabei, dass die Zahl der Aufgaben in diesen beiden Kompetenzbereichen höher ist als in den anderen Kompetenzbereichen, s. Blueprint). Die „Wissenschaftskom-

petenz“ (WI) ist mit einer Reliabilität von etwa 0.80-0.85 ebenfalls über die Zeit hinweg stabil erhoben. Bei der „Professionellen Handlungskompetenz“ (PH) und der „Kommunikativen Kompetenz“ (KO) ist in den beiden letzten Jahren ein Abfall unter 0.80 festzustellen, am deutlichsten bei der „Kommunikativen Kompetenz“ (KO) mit einer Reliabilität von 0.73 im Progresstest 2017 (siehe Abbildung 1, linke Abbildung).

Bei den Standardmessfehlern (siehe Abbildung 1, rechte Abbildung) ist eine deutliche Verringerung mit dem Jahr 2014 sichtbar, die auf die geringere Zahl der Aufgaben gegenüber dem SKPT 2013 zurückzuführen ist. Danach bleiben die Standardmessfehler für alle Kompetenzbereiche seit 2015 jeweils nahezu gleich. Auch hier sind die unterschiedlichen Zahlen von Aufgaben in den Kompetenzbereichen zu berücksichtigen, längere Skalen haben auch einen absolut größeren Standardmessfehler. Diese Reihenfolge bildet sich auch in der Abbildung ab, „Klinisch-praktische Kompetenz“ (KP) und „Klinisch-theoretische Kompetenz“ (KT) sind die längsten Skalen (im Blueprint jeweils 30 Aufgaben), die „Wissenschaftskompetenz“ (WI) ist mit lediglich 18 Aufgaben die kürzeste Skala.

In Tabelle 6 sind je Studienjahr und Kompetenzbereich die Mediane der Reliabilitäten in den Teilpopulationen mit gleicher Studiendauer aufgeführt. Zusätzlich zeigt Abbildung 2 die Verteilungen dieser jeweils 30 Einzelreliabilitäten getrennt für die Kompetenzbereiche. Während für die Kompetenzbereiche klinisch-praktische, klinisch-theoretische und Wissenschaftskompetenz (KP, KT bzw. WI) die Messzuverlässigkeiten in der Mehrzahl die Reliabilitäten über 0.7 liegen (KP 83%, WI 80%, KT 90%), sind diese Anteile bei der kommunikativen Kompetenz (KO) mit 37% und professionellem Handeln (PH) mit 47% deutlich niedriger. Insbesondere gibt es bei diesen einzelnen SKPTs und Studienjahren, für die niedrige Reliabilitäten unter 0.4 zu konstatieren sind (siehe Abbildung 2). Insbesondere scheinen in der Gruppe des 6. Studienjahres gehäuft niedrige Messzuverlässigkeiten aufzutreten (Anteile der Reliabilitäten über 0.7 für die Studienjahre 1-5:

Tabelle 5: Anzahl der Aufgaben (n_A), Reliabilität (Koeffizient glb) und Standardmessfehler (sem) der Kompetenzbereiche in den SKPTs der Jahre 2013–2017. Durch die Streichung einzelner Aufgaben im Post-Reviewverfahren sind die Anzahlen der Aufgaben je Bereich teilweise geringer als im Blueprint vorgesehen (siehe Tabelle 2).

Komp.- Bereich	Progresstest														
	2013			2014			2015			2016			2017		
	n_A	glb	sem	n_A	glb	sem	n_A	glb	sem	n_A	glb	sem	n_A	glb	sem
KO	26	0.85	1.75	21	0.83	1.42	22	0.80	1.71	21	0.78	1.66	21	0.73	1.47
KP	38	0.94	1.96	28	0.94	1.57	30	0.92	1.86	28	0.93	1.79	28	0.92	1.71
WI	18	0.84	1.54	17	0.83	1.30	17	0.84	1.46	18	0.85	1.48	17	0.81	1.43
PH	20	0.82	1.48	19	0.84	1.36	19	0.84	1.60	20	0.77	1.67	19	0.78	1.50
KT	38	0.92	1.92	29	0.95	1.54	29	0.92	1.78	29	0.91	1.78	27	0.92	1.73

Tabelle 6: Mediane der Reliabilitäten je Studienjahr und Kompetenzbereich aus den jeweils fünf SKPTs der Jahre 2013–2017.

Studienjahr	Kompetenzbereich				
	KO	KP	WI	PH	KT
1	0.779	0.852	0.783	0.799	0.816
2	0.754	0.782	0.833	0.748	0.751
3	0.658	0.789	0.782	0.721	0.791
4	0.568	0.732	0.767	0.657	0.790
5	0.635	0.751	0.799	0.539	0.772
6	0.436	0.752	0.748	0.452	0.736

KO 44%, KP 84%, WI 84%, PH 56%, KT 96%, siehe Tabelle 6).

3.2. Abgrenzung der Kompetenzbereiche („Diskriminanzvalidität“)

3.2.1. Abgrenzung der Kompetenzbereiche gegenüber jeweils allen anderen Bereichen

In Abbildung 3 und Abbildung 4 sind als Beispiele für die Progresstests der Jahre 2013 und 2017 die Ergebnisse der Zwei-Klassen-Diskriminanzanalysen eines Kompetenzbereichs jeweils gegenüber allen anderen Kompetenzbereichen dargestellt (die Ergebnisse der Jahre 2014 bis 2016 sind diesen sehr ähnlich). Auf der Ordinate sind die Werte der Diskriminanzfunktion der Einzelaufgaben als Boxplot abgetragen. Deutliche Überlagerungen der Boxplots zweier Kompetenzbereiche weisen auf Konstruktüberlappung bei den Aufgaben hin, bei deutlich separierten Boxen sind die Aufgabengruppen des zugehörigen Kompetenzbereichs von den anderen Aufgaben abzugrenzen.

So bilden die Aufgaben zur „Wissenschaftskompetenz“ (WI) in allen Jahren ein deutlich abgegrenztes Cluster, wer eine Aufgabe aus diesem Kompetenzbereich gelöst hat, hat auch mit höherer Wahrscheinlichkeit andere Aufgaben des Bereichs korrekt beantwortet (siehe Abbildung 3, Punkt c und Abbildung 4, Punkt c). Hingegen ist bei der „Klinisch-praktischen Kompetenz“ (KP) und „Klinisch-theoretischen Kompetenz“ (KT) eine deutliche Konstruktüberlappung festzustellen (siehe Abbildung 3, Punkte b, e und Abbildung 4, Punkte b, e).

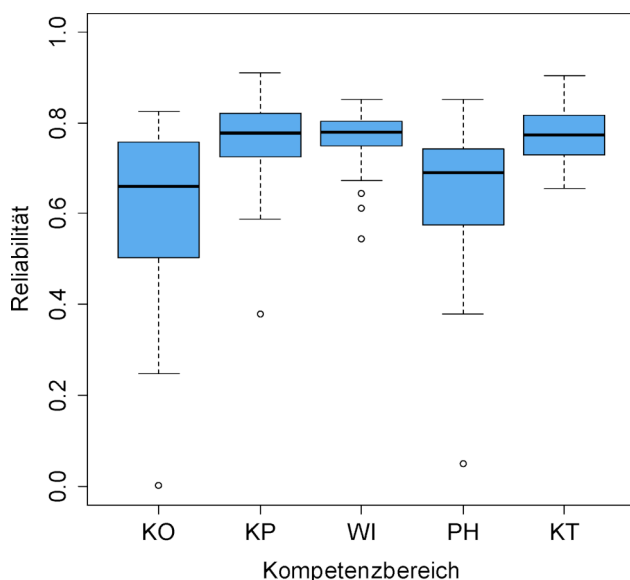


Abbildung 2: Reliabilitäten innerhalb der Studienjahre für die einzelnen Kompetenzbereiche. Jede „Box“ repräsentiert die 30 Einzelreliabilitäten der 5 Progresstests und 6 Studienjahre.

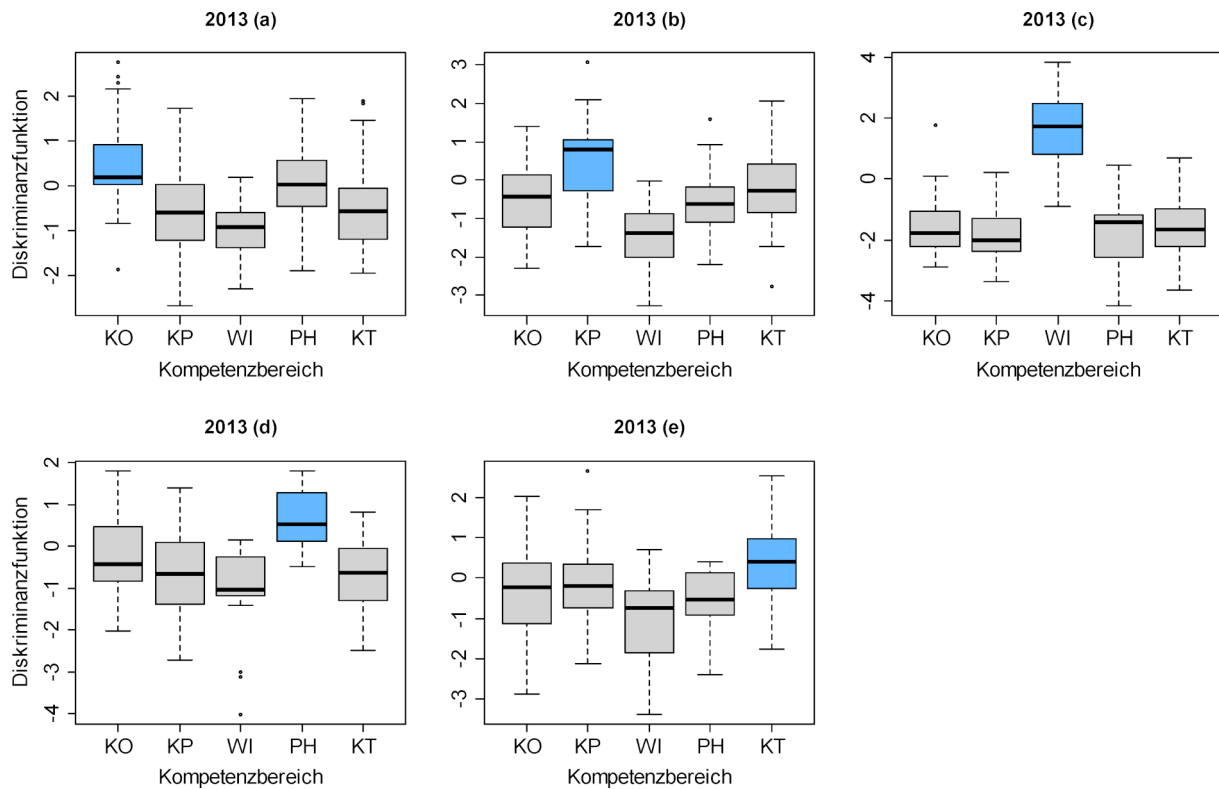


Abbildung 3: Verteilungen der Diskriminanzfunktionswerte der Aufgaben der Kompetenzbereiche nach Durchführung einer Zwei-Klassen Diskriminanzanalyse jedes Bereichs gegenüber allen anderen („one against the rest“) für den Progresstest 2013.

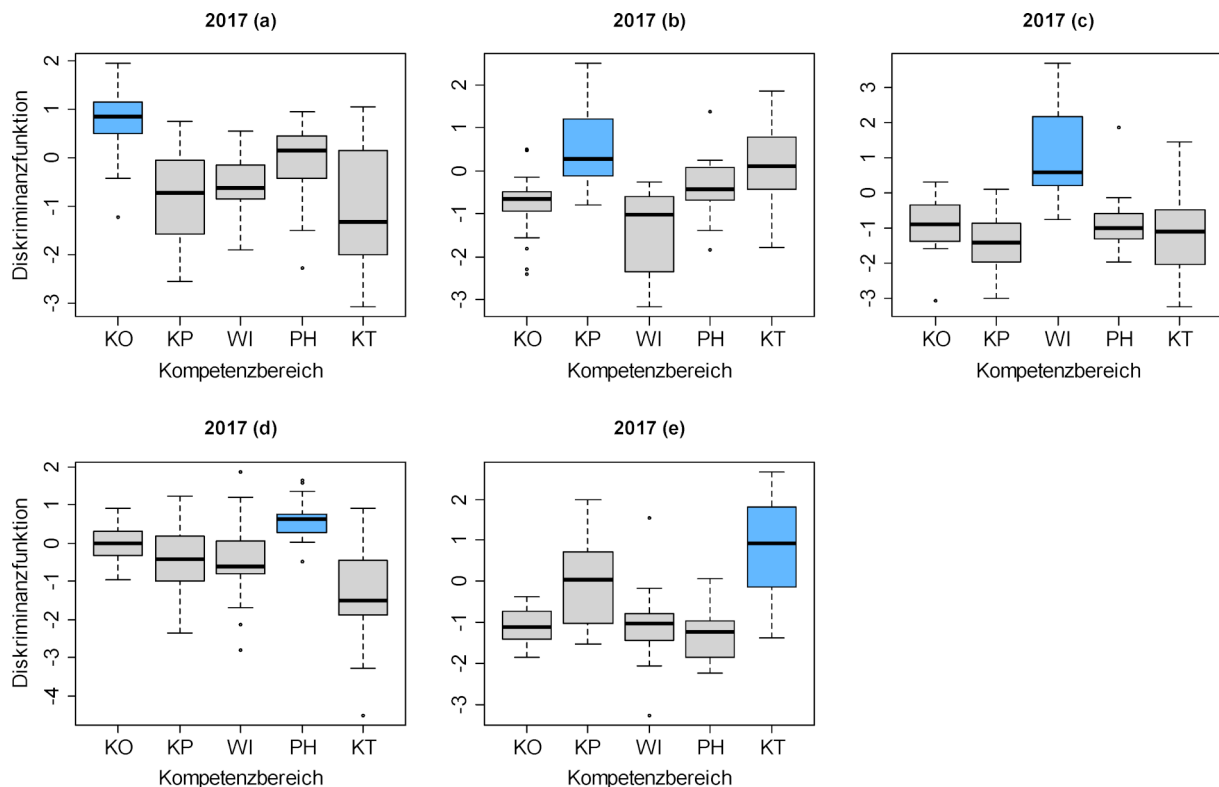


Abbildung 4: Verteilungen der Diskriminanzfunktionswerte der Aufgaben der Kompetenzbereiche nach Durchführung einer Zwei-Klassen Diskriminanzanalyse jedes Bereichs gegenüber alle anderen („one against the rest“) für den Progresstest 2017.

Tabelle 7: p-Werte der Tests auf Unterschied zwischen den Werten der Diskriminanzfunktion der Aufgaben eines Kompetenzbereichs der Progresstests 2013-2017 gegenüber den Aufgaben aller anderen Bereiche und kombinierter Gesamtttest nach Fisher (Fisher's combined probability test, FCPT).

Kompetenz-Bereich	Progresstest					FCPT
	2013	2014	2015	2016	2017	p
KO	0.055	0.017	<0.001	<0.001	0.002	<0.001
KP	<0.001	0.004	0.012	<0.001	0.026	<0.001
WI	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
PH	0.010	0.002	0.002	0.001	0.032	<0.001
KT	0.030	0.076	0.018	<0.001	<0.001	<0.001

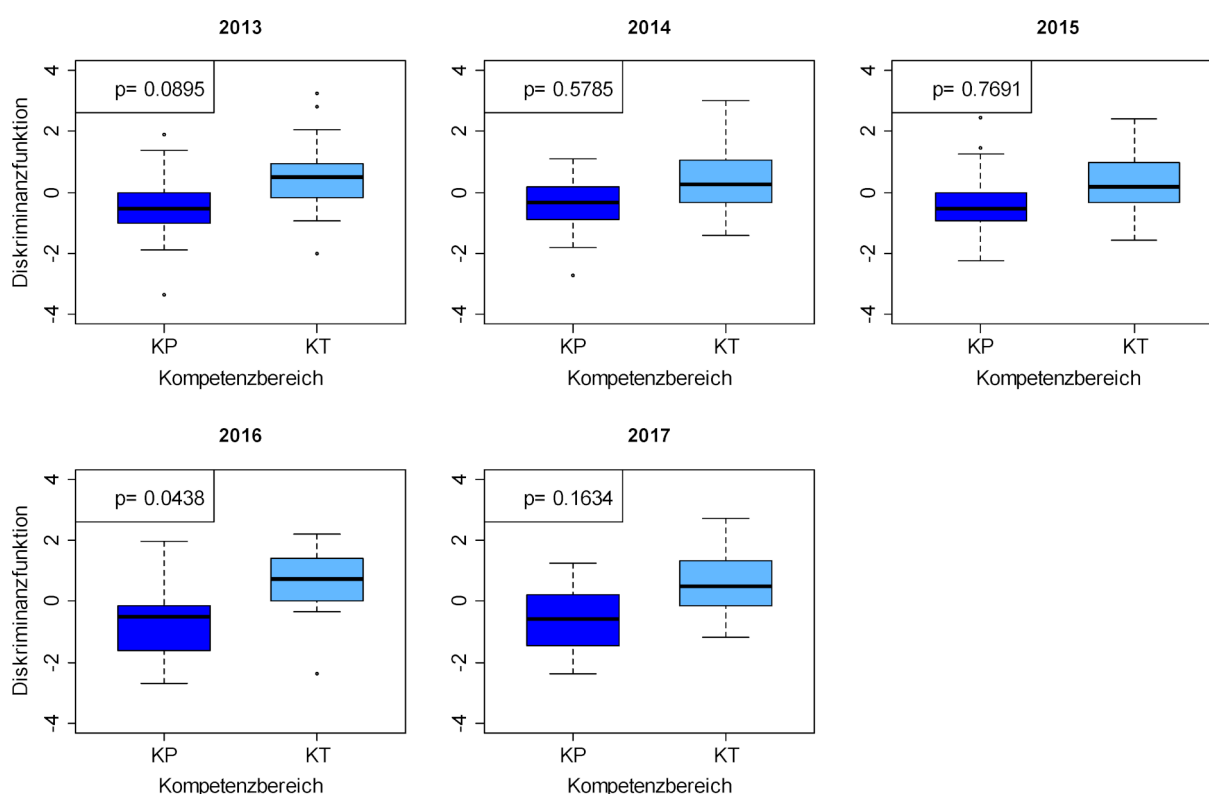


Abbildung 5: Verteilungen der Diskriminanzfunktionswerte der Aufgaben der Kompetenzbereiche „klinisch-praktisch“ und „klinisch-theoretisch“ nach Durchführung einer Zwei-Klassen-Diskriminanzanalyse für die Progresstests 2013–2017.

Zur inferenzstatistischen Testung, ob sich die einzelnen Kompetenzbereiche vom Gesamt aller anderen Bereiche unterscheiden, wurden die einzelnen p-Werte der Jahre 2013–2017 mit dem Gesamtttest nach Fisher kombiniert (man beachte, dass die einzelnen p-Werte nicht zu einer Signifikanztestung verwendet werden). Für alle fünf Bereiche ergeben sich signifikante Werte (siehe Tabelle 7).

3.2.2. Paarweise Abgrenzung der Kompetenzbereiche

Exemplarisch für die insgesamt 10 möglichen Paarvergleiche sind in Abbildung 5 die Werte der Diskriminanzfunktion der Aufgaben für den Vergleich der „Klinisch-praktischen Kompetenz“ (KP) mit der „Klinisch-theoretischen Kompetenz“ (KT) sowie in Abbildung 6 die der „Wissenschaftskompetenz“ (WI) mit der „Klinisch-theoretischen Kompetenz“ (KT) dargestellt.

Die aus den Einzelwerten der verschiedenen Durchgänge des SKPT kombinierten Signifikanzwerte aller Paarvergleiche sind in Tabelle 8 aufgeführt. Mit Ausnahme des klinisch-praktischen (KP) und des klinisch-theoretischen Kompetenzbereichs (KT) sind alle Paarvergleiche signifikant.

Die Signifikanzen bleiben auch nach Bonferroni-Holm-Adjustierung der zehn Tests erhalten: Hierzu werden die einzelnen p-Werte der Tests ihrer Größe nach aufsteigend angeordnet ($p_{[1]} \leq p_{[2]} \leq \dots \leq p_{[10]}$). Als signifikant gelten genau die $p_{[k]}$, bei denen für alle $i \leq k$ die Ungleichungen $p_{[i]} \leq \alpha / (11 - i)$ erfüllt sind. Ist nur für ein einziges kleineres $p_{[i]}$ die Ungleichung nicht erfüllt, kann kein größeres $p_{[k]}$ noch als signifikant gelten.

Bei acht der Tests ist $p < 0.001$, somit sind diese wegen $p_{[1]} \leq 0.05/10$, $p_{[2]} \leq 0.05/9$, ... $p_{[8]} \leq 0.05/3$ signifikant, weiter ist $p_{[9]} = 0.002 \leq 0.05/2 = 0.025$ ebenfalls signifikant. Ledig-

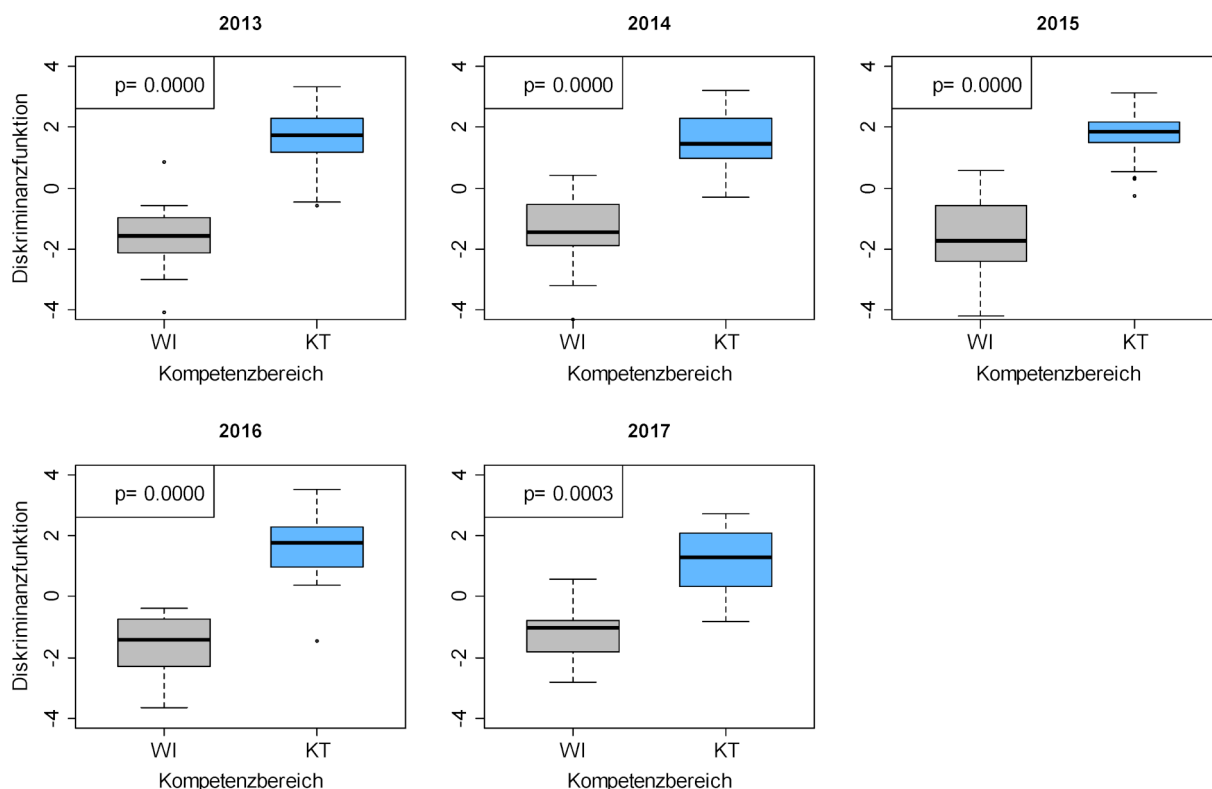


Abbildung 6: Verteilungen der Diskriminanzfunktionswerte der Aufgaben der Kompetenzbereiche „Wissenschaftskompetenz“ und „klinisch-theoretische Kompetenz“ nach Durchführung einer Zwei-Klassen-Diskriminanzanalyse für die Progresstests 2013–2017.

Tabelle 8: Kombinierte Signifikanztests nach Fisher (FCPT) der paarweisen Tests auf Unterschied der Werte der Diskriminanzfunktion zwischen den Aufgaben zweier Kompetenzbereiche.

	Kompetenzbereich				
	KO	KP	WI	PH	KT
KO	–				
KP	<0.001	–			
WI	<0.001	<0.001	–		
PH	0.002	<0.001	<0.001	–	
KT	<0.001	0.091	<0.001	<0.001	–

lich der dem Vergleich der Kompetenzbereiche KP und KT zugeordnete p-Wert $p_{[10]}=0.091$ erfüllt nicht die Bedingung $p_{[10]}=0.05/1$.

4. Zusammenfassung und Diskussion

Die Kompetenzbereiche „Klinisch-praktische Kompetenz“ (KP), „Klinisch-theoretische Kompetenz“ (KT) und „Wissenschaftskompetenz“ (WI) wurden mit den fünf SKPTs über die Jahre hinweg reliabel erfasst (Reliabilität über 0.80) Nicht ganz so zufriedenstellend sind die Aufgaben zu den Bereichen „Kommunikative Kompetenz“ (KO) und „Professionelle Handlungskompetenz“ (PH) mit Reliabilitäten noch über 0.73. Die Aufgaben zu den verschiedenen Kompetenzbereichen repräsentieren auch empirisch

unterschiedliche Bereiche. Eine Ausnahme bilden die Aufgabengruppen zum klinisch-praktischen und klinisch-theoretischen Kompetenzbereich, für die keine Abgrenzung voneinander empirisch nachzuweisen ist ($p=0.091$). Die Messzuverlässigkeiten innerhalb der Jahrgangskohorten weisen, von wenigen Ausnahmen abgesehen, für die Bereiche klinisch-praktische, klinisch-theoretische und Wissenschaftskompetenz (KP, KT, WI) zufriedenstellende Werte auf. Einschränkungen sind wie oben bei den Bereichen Kommunikation und professionellem Handeln (KO, PH) festzustellen, bei denen die Reliabilitäten doch häufig 0.7 nicht erreicht. Dabei muss jedoch berücksichtigt werden, dass diese Bereiche lediglich mit 22 bzw. 20 Aufgaben abgedeckt werden, was auch bei gut vorbereiteten summativen Fachprüfungen an den Fakultäten im Allgemeinen nicht ausreicht, mit Fragen des Typs A eine hinreichend hohe Reliabilität zu erreichen.

Damit ist gezeigt, dass in kompetenzorientierten Wissenstests die Erstellung von Aufgaben zur Erfassung von handlungsrelevantem Wissen auch bei begrenzter Zahl von Aufgaben möglich ist und – in der Terminologie von Cronbach und Meehl [2] – die im Blueprint intendierten unterschiedlichen Konstrukte für die Kompetenzbereiche abbilden.

Ausnahme hiervon sind der klinisch-praktische und der klinisch-theoretische Kompetenzbereich (KP/KT), die sich zwar signifikant von den anderen Kompetenzbereichen, nicht aber untereinander unterscheiden. Die Tatsache, dass die beiden Bereiche nicht voneinander abzugrenzen sind, kann auf Grund des verwendeten Verfahrens der Diskriminanzanalyse, das explizit ein „construct overlap“ erlaubt, nicht als Erklärung dienen. Grund für die Nicht-separierbarkeit könnte eine zu wenig eindeutige Vorgabe für die Klassierung der Aufgaben in die beiden Bereiche sein. Eine alternative Erklärungsmöglichkeit bestünde darin, dass diese im Sinn einer fachlichen Kompetenz stärker inhaltlich zusammenhängen und dass bei den teilnehmenden Studierenden die Kenntnisse in beiden Bereichen zum großen Teil zeitlich parallel erworben werden. Dies würde dazu führen, dass trotz semantischer Verschiedenheit der beiden Bereiche, kein empirischer Unterschied bei den Lösungsmustern nachweisbar ist: wer bei den Aufgaben eines Bereichs gute Kenntnisse besitzt, besitzt sie auch bei denen des anderen.

Methodisch hat sich die Diskriminanzanalyse der Hauptkomponenten (DAPC) als geeignet erwiesen, die zugrundeliegende Struktur der Achse „Kompetenzbereiche“ des Blueprints empirisch nachzuzeichnen.

Als Konsequenz aus den Ergebnissen sollte für zukünftige Progresstests eine Verbesserung der Messzuverlässigkeit der Kompetenzbereiche „kommunikative“ (KO) und „professionelle Handlungskompetenz“ (PH) angestrebt werden. Dies könnte etwa mit einer Erhöhung der zugehörigen Zahl von Aufgaben erreicht werden. Um den Umfang des SKPT nicht zu vergrößern, wäre eine Verminderung der Aufgabenzahl für den „klinisch-praktischen“ (KP) und „klinisch-theoretischen“ (KT) Bereich denkbar. Die inhaltliche Abgrenzung dieser beiden Bereiche sollte ebenfalls geprüft werden. Lässt sich auch hier keine empirische Trennung erreichen, könnten diese für die Rückmeldung an die teilnehmenden Studierenden auch zusammengefasst werden.

Anmerkungen

1. Teile des Inhalts der Studie für die Progresstests 2013 und 2014 wurden auf der Tagung der GMA 2014 in Hamburg und der RIME 2015 in München berichtet [14], [15].
2. Eine Originalquelle zur Diskriminanzanalyse von Hauptkomponenten ist den Autoren nicht bekannt, die älteste von uns gefundene Erwähnung findet sich in einer Arbeit zur Verteilung großer Säugetiere in einem Nationalpark in Tansania [16].

Förderung

Die Arbeit entstand im Rahmen des vom Bundesministerium für Bildung und Forschung geförderten Projekts MERLIN II (01PL17011C).

Interessenkonflikt

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

Anhänge

Verfügbar unter

<https://www.egms.de/de/journals/zma/2020-37/zma001299.shtml>

1. Anhang_1.pdf (391 KB)
Beispiele für Fragen aus den Kompetenzbereichen des studentischen kompetenzorientierten Progresstests 2017

Literatur

1. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull.* 1959;56(2):81-105. DOI: 10.1037/h0046016
2. Cronbach LJ, Meehl P E. Construct validity in psychological tests. *Psychol Bull.* 1955;52(4):281-302. DOI: 10.1037/h0040957
3. Frank JR, Snell LS, Cate OT, Holmboe ES, Carraccio C, Swing SR, Harris P, Glasgow NJ, Campbell C, Dath D, Harden RM, Iobst W, Long DM, Mungroo R, Richardson DL, Sherbino J, Silver I, Taber S, Talbot M, Harris KA. Competency-based medical education: theory to practice. *Med Teach.* 2010;32(8):638-645. DOI: 10.3109/0142159X.2010.501190
4. Hahn EG, Fischer MR. Nationaler Kompetenzbasierter Lernzielkatalog Medizin (NKLM) für Deutschland: Zusammenarbeit der Gesellschaft für Medizinische Ausbildung (GMA) und des Medizinischen Fakultätentages (MFT). *GMS Z Med Ausbild.* 2009;26(3):Doc35. DOI: 10.3205/zma000627
5. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency based medical education. *Med Teach.* 2010;32(8):676-682. DOI: 10.3109/0142159X.2010.500704
6. Hughes DJ. Psychometric validity: Establishing the accuracy and appropriateness of psychometric measures. In: Irwing P, Booth T, Hughes DJ, editors. *Wiley handbook of psychometric testing: A multidisciplinary approach to survey, scale, and test development.* Hoboken, NJ: Wiley; 2018. DOI: 10.1002/9781118489772.ch24
7. Jackson, P, Agunwamba, C. Lower bounds for the reliability of the total score on a test composed of non-homogeneous items I: Algebraic lower bounds. *Psychomet.* 1977;42(4):567-578. DOI: 10.1007/BF02295979
8. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Gen.* 2010;11:94. DOI: 10.1186/1471-2156-11-94
9. Lord FM, Novick MR. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley; 1968.

10. Lotte F, Bougrain L, Cichocki A, Clerc M, Congedo M, Rakotomamonjy A, Yger F. A Review of Classification Algorithms for EEG-based Brain-Computer Interfaces: A 10-year Update. *J Neural Eng.* 2018;15(3):031005. DOI: 10.1088/1741-2552/aab2f2
11. Lurie, SJ. History and practice of competency-based assessment. *Med Educ.* 2012;46(1):49-57. DOI: 10.1111/j.1365-2923.2011.04142.x
12. McGill DA, van der Vleuten CP, Clarke MJ. A critical evaluation of the validity and the reliability of global competency constructs for supervisor assessment of junior medical trainees. *Adv Health Sci Educ Theory Pract.* 2013;18(4):701-725. DOI: 10.1007/s10459-012-9410-z
13. McHarg J, Bradley P, Chamberlain S, Ricketts C, Searle J, McLachlan JC. Assessment of progress tests. *Med Educ.* 2005;39(2):221-227. DOI: 10.1111/j.1365-2929.2004.02060.x
14. Möltner A, Wagener S, Jünger J. Empirische Struktur des kompetenzbasierten studentischen Progresstests: Reliabilität und diskriminante Validität von Kompetenzbereichen. In: *Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA). Hamburg, 25.-27.09.2014. Düsseldorf: German Medical Science GMS Publishing House; 2014. DocV443. DOI: 10.3205/14gma308*
15. Möltner A, Wagener S, Timbil S, Gornostayeva M, Jünger J. Empirical Structure of a Competency-Based Progress Test. In: *4th Research in Medical Education (RIME) Symposium 2015. München, 19.-21.03.2015. Düsseldorf: German Medical Science GMS Publishing House; 2015. DocS1B3. DOI: 10.3205/15rime14*
16. Morton-Griffiths M. The numbers and distribution of large mammals in Ruaha National Park, Tanzania. *E Afr Wildl J.* 1975;13:121-140. DOI: 10.1111/j.1365-2028.1975.tb00127.x
17. Newton PE, Shaw SD. Validity in educational and psychological assessment. Los Angeles: Sage; 2014. DOI: 10.4135/9781446288856
18. Nouns ZM, Georg W. Progress testing in German speaking countries. *Med Teach.* 2010;32(6):467-470. DOI: 10.3109/0142159X.2010.485656
19. Onatski A. Testing hypotheses about the number of factors in large factor models. *Econometr.* 2009;77(5):1447-1479. DOI: 10.3982/ECTA6964
20. Patterson F, Zibarras L, & Ashworth V. Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. *Med Teach.* 2016;38(1):3-17. DOI: 10.3109/0142159X.2015.1072619
21. Raïche G, Walls TA, Magis D, Riopel M, Blais JG. Non-graphical solutions for Cattell's scree test. *Methodol.* 2013;9(1):23-29. DOI: 10.1027/1614-2241/a000051
22. Ravesloot CJ, Van der Schaaf MF, Muijtjens AMM, Haaring C, Kruitwagen CL, Beek FJ, Bakker J, Van Schaik JP, Ten Cate TJ. The don't know option in progress testing. *Adv Health Sci Educ.* 2015;20(5):1325-1338. DOI: 10.1007/s10459-015-9604-2
23. Schindler C, Bauer J, Strasser A, Schlomske-Bodenstein N, Seidel T, Prenzel M. Prüfungen als Indikator für den Studienerfolg. In: Berthold C, Jorzik B, Meyer-Guckel V, editors. *Handbuch Studienerfolg.* Essen: Edition Stifterverband; 2015. p.62-79.
24. Schuwirth LW, van der Vleuten CP. The use of progress testing. *Perspect Med Educ.* 2012;1(1):24-30. DOI: 10.1007/s40037-012-0007-2
25. Sijsma K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika.* 2009;74(1):107-120. DOI: 10.1007/s11336-008-9101-0
26. Steinhäuser J, Chenot JF, Roos M, Ledig T, Joos S. Competence-based curriculum development for general practice in Germany: a stepwise peer-based approach instead of reinventing the wheel. *BMC Res Notes.* 2013;6(1):314. DOI: 10.1186/1756-0500-6-314
27. Tax DM, Duin RP. Using two-class classifiers for multiclass classification. In: *Proceedings of the 16th International Conference on Pattern Recognitions. Vol. 2; 2002 Aug 11-15; Quebec, Kanada. IEEE Press; 2002. p.124-127. DOI:10.1109/ICPR.2002.1048253*
28. Van der Vleuten CP, Verwijnen GM, Wijnen WH. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Med Teach.* 1996;18(2):103-109. DOI: 10.3109/01421599609034142
29. Wagener S, Fleig A, Möltner A. Warum sind im studentischen kompetenzorientierten Progresstest in jedem Studienjahr etwa 25% der Antworten falsch? In: *Gemeinsame Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA) und des Arbeitskreises zur Weiterentwicklung der Lehre in der Zahnmedizin (AKWLZ). Münster, 20.-23.09.2017. Düsseldorf: German Medical Science GMS Publishing House; 2017. Doc157. DOI: 10.3205/17gma157*
30. Wagener S, Gaitzsch E, Brass K, Heid J, Herrmann L, Günther J, Ney S, Müller A, Dikmen HO, Zimmermann K, Yilmaz OK, Wittstock F, Alhalabi O, Park J, Harapan BN, Kollmeier B, Ronellenfitch L, Mayer J, Baumann T, Daunert L, Brüstle P, Mohr D, Schüttelpelz-Brauns K, Van der Beken A, Jünger J, Möltner A. Videofragen im studentischen kompetenzorientierten Progresstest. In: *Gemeinsame Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA) und des Arbeitskreises zur Weiterentwicklung der Lehre in der Zahnmedizin (AKWLZ). Münster, 20.-23.09.2017. Düsseldorf: German Medical Science GMS Publishing House; 2017. Doc110. DOI: 10.3205/17gma110*
31. Wagener S, Möltner A, Fleig A, Feistner L, Heid J, Brass K, Holz T, Weber M, Pflaum P, Rogg D, Kellermann F, Berg L, Breithaupt MH, Dehmel L, Grad A, Xiang Jin J, Hai-Ning Lu K, Müller A, Rinawi T, Shang V, Zimmermann K, Alhalabi O, Park J, Grupp M, Klauth A, Lepper A, Lichnock Z, Mayer J, Hollmann A, Meuth C, Siegel F, Peitz N, Brüstle P, Mohr D, Schüttelpelz-Brauns K, Würth G, Jünger J, Burkert M. "Was will ich prüfen?" - Erweiterung des Blueprints im studentischen kompetenzorientierten Progresstest. In: *Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA). Wien, 19.-22.09.2018. Düsseldorf: German Medical Science GMS Publishing House; 2018. Doc19.5. DOI: 10.3205/18gma338*
32. Wagener S, Möltner A, Timbil S, Fleig A, Feistner L, Heid J, Brass K, Burkert M. "Da bin ich mir sicher" - Confidence rating im studentischen kompetenzorientierten Progresstest [Bericht über Forschungsergebnisse]. In: *Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA). Wien, 19.-22.09.2018. Düsseldorf: German Medical Science GMS Publishing House; 2018. Doc15.2. DOI: 10.3205/18gma067*
33. Wagener S, Möltner A, Timbil S, Gornostayeva M, Schultz JH, Brüstle P, Mohr D, Van der Beken A, Better J, Fries M, Gottschalk M, Günther J, Herrmann L, Kreisel C, Moczek T, Illg C, Jassowicz A, Müller A, Niesert M, Strübing F, Jünger J. Development of a competencybased formative progress test with student-generated MCQs: Results from a multi-centre pilot study. *GMS Z Med Ausbild.* 2015;32(4):Doc46. DOI: 10.3205/zma000988
34. Wrigley W, van der Vleuten CP, Freeman A, Muijtjens A. A systemic framework for the progress test: strengths, constraints and issues: AMEE Guide No. 71. *Med Teach.* 2012;34(9):683-697. DOI: 10.3109/0142159X.2012.704437
35. Zmud RW, Sampson JP, Reardon RC, Lenz JG, Byrd TA. Confounding Effects of Construct Overlap: An Example from IS User Satisfaction Theory. *Inform Technol People.* 1994;7(2):29-45. DOI: 10.1108/09593849410074061

Korrespondenzadresse:

Dr. phil. Andreas Möltner
Medizinische Fakultät Heidelberg, Kompetenzzentrum
für Prüfungen in der Medizin Baden-Württemberg, Im
Neuenheimer Feld 346, 69120 Heidelberg, Deutschland
andreas.moeltner@med.uni-heidelberg.de

Bitte zitieren als

Möltner A, Wagener S, Burkert M. Measuring competency-relevant
knowledge in the competency-oriented student progress test. *GMS J
Med Educ.* 2020;37(1):Doc6.
DOI: 10.3205/zma001299, URN: urn:nbn:de:0183-zma0012993

Artikel online frei zugänglich unter

<https://www.egms.de/en/journals/zma/2020-37/zma001299.shtml>

Eingereicht: 04.02.2019

Überarbeitet: 01.08.2019

Angenommen: 14.10.2019

Veröffentlicht: 17.02.2020

Copyright

©2020 Möltner et al. Dieser Artikel ist ein Open-Access-Artikel und
steht unter den Lizenzbedingungen der Creative Commons Attribution
4.0 License (Namensnennung). Lizenz-Angaben siehe
<http://creativecommons.org/licenses/by/4.0/>.