Increase in medical knowledge during the final year of undergraduate medical education in Germany

Abstract

Aims: In Germany, the final year of undergraduate medical education ('practice year') consists of three 16-week clinical attachments, two of which are internal medicine and surgery. Students can choose a specific specialty for their third 16-week attachment. Practice year students do not receive specific teaching to prepare them for the National Licensing Examination. It is unknown whether knowledge levels increase during this year. This study aimed at assessing knowledge at the beginning and the end of the final year of medical school.

Methods: Three hundred pre-selected United States Medical Licensing Examination type items from ten medical disciplines were reviewed by ten recent medical graduates from the Netherlands and Germany. The resulting test included 150 items and was taken by 77 and 79 final year medical students from Göttingen and Hamburg at the beginning and the end of their practice year, respectively.

Results: Cronbach's α of the pre- and post-test was 0.75 and 0.68, respectively. Mean percent scores in the pre- and post-test were 63.9±6.9 and 69.4±5.7, respectively ($p<0.001$; effect size calculated as Cohen's d: 0.87). In individual students, post-test scores were particularly high for items related to their specific chosen specialty.

Conclusion: The knowledge test used in this study provides a suitable external tool to assess progress of undergraduate medical students in their knowledge during the practice year. The pre-test may be used to guide individual learning behaviour during this final year of undergraduate education.

Keywords: undergraduate medical education, practice, knowledge

Tobias Raupach[1]
Daniela Vogel[2]
Sarah Schiekirka[1,3]
Carolina Keijsers[4]
Olle Ten Cate[5]
Sigrid Harendza[2]

1 University Medical Centre Göttingen, Department of Cardiology and Pneumology, Göttingen, Germany

2 University Medical Centre Hamburg-Eppendorf, III. Department of Internal Medicine, Hamburg, Germany

3 University Medical Centre Göttingen, Study Deanery, Göttingen, Germany

4 University Medical Centre Utrecht, Department of Geriatric Medicine, Utrecht, the Netherlands

5 University Medical Centre Utrecht, Center for Research and Development of Education, Utrecht, the Netherlands

Introduction

Medical students in Germany spend their final year of undergraduate medical studies, the so-called "practice year", full time in a clinical environment. While all students are required to spend 16 weeks each in internal medicine and general surgery, they can choose one additional specialty rotation for the remaining 16 weeks of their practice year. Until 2012, the practice year was followed by a final National Licensing Examination (NLE) comprising 320 multiple choice questions which cover both basic and clinical knowledge. Medical schools did not routinely offer specific training for this high-stakes examination. As a consequence, German students had to find a balance between maximizing their learning gain regarding clinical skills and meeting the assessment requirements regarding basic and clinical knowledge during the final year. In the absence of formal assessments during this year, students were unable to monitor their progress.

Progress testing is designed to foster learning processes that are characterized as meaning oriented in contrast to reproduction oriented learning [1]. Several countries have developed progress tests to measure general medical knowledge progression of their medical students [2], [3], [4]. Furthermore, some students use the feedback provided by progress testing to guide their own learning activities [5]. Thus, a formative test of knowledge for final year students at the beginning and end of the practice year might be a helpful tool to provide feedback regarding knowledge gain during the final year. This study uses a 150 item multiple-choice knowledge test which had ori-

ginally been designed as a suitably balanced and non-biased test for the comparison of final year students from different European countries [6]. For the purpose of the present study, the test was administered at two German universities. The primary aim of this study was to assess the increase in overall knowledge as assessed with this test during the practice year. We hypothesized that exam scores would be significantly higher at the end of the practice year as exposure to a clinical environment might also foster the acquisition of factual knowledge. We expected this effect to be most pronounced in those subject areas that students specifically addressed during their 16-week specialty rotation.

Methods

Use of multiple-choice questions

It has been shown that multiple-choice questions (MCQs) are a valid method of assessing cognitive knowledge [7]. MCQ examinations can yield high reliability which is an important requirement to be able to distinguish between groups or individuals [8]. Furthermore, well-constructed MCQs can also assess higher-order cognitive processes such as interpretation and application of knowledge skills rather than just testing the recall of facts [9]. Different formats of MCQs have been designed, including variations on the number of item branches. The United States Medical Licensing Examinations (USMLE) typically include single best answer questions with five branches, and the same format is used in the German NLE.

Selection of examination items

We used a 150-item knowledge test that was compiled from 1000 -freely available United States Medical Licensing Examination Step 2 type items [10]. Initially, 300 items had been selected and adapted according to the following criteria:

1. All items had to belong to major medical disciplines: general medicine, anaesthesiology & emergency medicine, internal medicine, surgery, urology, obstetrics and gynaecology, paediatrics, neurology, psychiatry, clinical pharmacology. Small disciplines like ear-nose-and-throat or dermatology were excluded.
2. All items were based on patient cases.
3. Diseases specific to the Americas (e.g. Rocky Mountain spotted fever) were not included.
4. Items including figures of any sort (e.g. X-rays or ECGs) were excluded for copyright reasons.
5. Answers for each question all had to belong to the same category (e.g. diagnostics, therapy or other). Items including, for instance, two diagnostic and three therapeutic answers were not chosen, or adapted to fit this criterion.
6. If different items covered the same topic or disease, only one item was included.

In February 2011, the 300 selected items plus answers were reviewed by five recent medical graduates from both Hamburg and Utrecht University in the Netherlands (total n=10) to identify items which seemed appropriate in more than one European country. Every rater was asked to choose 50% of the items to match best the expected content level of graduates in their country, with a fixed number per discipline to ensure content validity. The final test included 150 items chosen by at least 2 of the 5 raters from each country. A suitability score was calculated eventually for the final test by checking for difficult versus easy items and basic science versus clinical knowledge items by one rater per country resulting in a well balanced test with intraclass correlations of 0.85 and 0.71, respectively. A pilot test with a total of 56 students from Germany and the Netherlands yielded a Cronbach's alpha for internal consistency of 0.79 [6].

Application of the formative pre- and post-tests

A German version of this newly developed test was used in our project as a formative pre- and post-test for final year undergraduate medical students in Germany, offered to 286 students (164 students in Göttingen and 122 students in Hamburg) at the beginning of their practice year (April 2011). Students received an e-mail outlining the study rationale and aims and were invited to participate in the pre-test. Students were also informed that they would be released from their clinical duties on the day of the test; this had been agreed upon with all teaching hospitals at both study sites.
A second test featuring the same items as the first one was offered to the same student groups at the end of their final year, just before students started to prepare for their NLE in spring 2012. Questions and answers were not made available to study participants and all papers were collected after writing the test. However, students were informed about their results via e-mail. Socio-economic data and information on study time spent abroad and choice of specialty rotation of participating students were also recorded, and participation was voluntary. This study was approved by the Hamburg State Ethics Committee.

Data collection and statistical analysis

Data collected on questionnaires and examination papers were manually transferred to the statistical software package SPSS 19.0 (SPSS Inc., Chicago, Illinois, USA). Differences between the two student cohorts participating in the pre- and the post-test, respectively, were assessed using χ^2-tests (dichotomous variables) and t-tests (continuous variables). Effect sizes were calculated as Cohen's d with values of 0.2 indicating small and values of 0.8 indicating large effects [11]. Item characteristics of the pre- and post-examination were assessed in terms of item difficulty, corrected item-total correlations and Cronbach's α as a measure of internal consistency. In order to detect a difference in exam percent scores of 3% (e.g., 68%

versus 65% with standard deviations of 6.5% in each group) at a significance level of 5% with a statistical power of 80%, a minimum of 58 students had to be enrolled to participate in both the pre- and the post-test (equivalent to a longitudinal response rate of 20%). We chose a 3% difference as important as it usually represents approximately one-third of a step in the German marking system. Data are presented as mean±standard deviation or percentages (n), as appropriate. Significance levels were set to $p<0.05$.

Results

Response rate and subject characteristics

The pre-test was taken by 77 students, and the post-test was taken by 79 students (response rates 26.9% and 27.6%, respectively). The proportion of female students was 66.2% in the pre- and 73.4% in the post-test, respectively. Response rates were higher in Hamburg and differed slightly between the pre- and the post-test at both study sites (Hamburg: 45/122 (36.9%) versus 58/122 (47.5%); Göttingen: 32/164 (19.5%) versus 21/164 (12.8%)). A total of 47 students took both the pre- and the post-test. Subject characteristics of the two cohorts are displayed in Table 1. As expected, students taking the post-test were significantly older than students taking the pre-test. There were no significant differences between the cohorts regarding gender, mother tongue, previous vocational training, spending parts of the practice year abroad and choice of the specialty rotation.

Item analysis

Cronbach's α of the pre- and post-test was 0.75 and 0.68, respectively. Item difficulty ranged from 0.03 to 1.00 (mean 0.64) in the pre- and from 0.04 to 1.00 (mean 0.69) in the post-test. The percentage of items with a difficulty between 0.4 and 0.8 was 56.7% (n=85) and 50.7% (n=76) in the pre- and post-test, respectively. Corrected item-total correlations of exam items ranged from -0.20 to 0.39 (mean 0.13) in the pre- and from -0.32 to 0.45 (mean 0.10) in the post-test. The percentage of items with positive discriminatory power was 85.3% (n=128) and 70.7% (n=106) in the pre- and post-test, respectively.

Student performance

Students achieved a mean percent score of 63.9±6.9 in the pre- and 69.4±5.7 in the post-test. ($T(154)=-5.376$; $p<0.001$; t-test for independent samples). The effect size of this difference calculated as Cohen's d was 0.87, indicating a large effect. A test for dependent samples in the subset of 47 students who took both tests yielded a similar result: 64.6±6.7 in the pre- and 69.6±5.3 in the post-test ($T(46)=-7.299$; $p<0.001$). Analysis of exam results for specific specialties (see Table 2) revealed that students who had chosen anaesthesiology & medical emergencies as their specialty rotation performed no better in items related to anaesthesiology & medical emergencies in the post-test than students who had chosen any other specialty (11.7±1.4 versus 11.7±1.6 out of 15 points; $p=0.985$), and a similar pattern of results was observed for neurology (9.9±2.0 vs. 10.2±1.8 out of 15 points; $p=0.639$). However, students who had chosen paediatrics achieved higher post-test scores in items related to paediatrics than students who had chosen any other specialty (13.6±1.3 vs. 11.6±1.5 out of 15; $p<0.001$; Cohen's $d=1.40$), and the same was true for obstetrics and gynaecology (7.8±1.3 vs. 6.7±1.4 out of 10 points; $p=0.017$; Cohen's $d=0.81$). Anecdotally, five out of the 14 students who had chosen paediatrics as their specialty rotation answered all paediatrics items correctly. In contrast, only one out of the 65 students who had chosen a specialty other than paediatrics answered all paediatrics items correctly.

Discussion

Using a 150-item test consisting of USMLE type items as a formative pre- and post-test for final year students in Germany, this study demonstrates a significant increase in knowledge levels after the practice year of undergraduate medical education. This increase appeared to be greatest regarding items related to students' specialty choices. A critical appraisal of individual performance levels by means of a formative test might be helpful to guide students' self-study during this year. To this date, no validated knowledge test has been available for this purpose. Our newly developed test might close this gap by providing a tool for formative assessment of medical students in their final year. In fact, individual feedback from several study participants indicated that they appreciated being able to assess their own knowledge under simulated 'exam conditions' and that they used their pre-test results to guide individual learning during the practice year.

Increase in exam performance levels during the practice year

Overall, performance levels as assessed in the post-test were rather low, and there are a number of potential explanations for this finding. Given that we used a formative rather than a summative exam, students might not have been sufficiently incentivized to achieve the maximum scores they would have been able to score [12]. On the other hand, it might be hypothesized that students participating in the study were highly motivated to know more about their performance level and would thus have tried the best they could to answer the exam items. However, they also might not have been used to the wording in the USMLE type item format.

Even taking into account these potential limitations, the increase from the pre- to the post-test results observed

Table 1: Characteristics of study participants. P values were derived from independent t-tests (age) and χ^2-tests (all other variables), respectively.

	Pre-Test (n = 77)	Post-Test (n = 79)	p value	Pre- and Post-Test (n = 47)
Age, mean ± SD	26.5 ± 2.1	27.3 ± 3.0	**0.048**	26.4 ± 2.3
Percentage of female students, % (n)	66.2 (51)	73.4 (58)	0.328	74.5 (35)
Percentage of students who are German native speakers, % (n)	84.4 (65)	88.6 (70)	0.443	91.5 (43)
Percentage of students who had taken vocational training before studying medicine, % (n)	16.4 (10)	17.7 (14)	0.836	19.1 (9)
Percentage of students who spent part of their practical year abroad, % (n)	21.3 (13)	30.4 (24)	0.228	21.3 (10)
Specialty choice for the 16-week specialty rotation				
Specialty module Anaesthesiology & emergency medicine, % (n)	19.7 (12)	17.7 (14)	0.620	14.9 (7)
Specialty module Paediatrics, % (n)	19.7 (12)	17.7 (14)		19.1 (9)
Specialty module Obstetrics & gynaecology, % (n)	16.4 (10)	13.9 (11)		14.9 (7)
Specialty module Neurology, % (n)	11.5 (7)	11.4 (9)		14.9 (7)
Specialty module Ear-Nose-Throat, % (n)	3.3 (2)	8.9 (7)		4.3 (2)
Specialty module Urology, % (n)	0.0 (0)	5.1 (4)		0 (0)
Specialty module Radiology, % (n)	4.9 (3)	5.1 (4)		6.4 (3)
Specialty module Other, % (n)	24.6 (15)	20.3 (16)		25.5 (12)

Table 2: Student performance in the pre- and post-test by specialty. Results are given as percentages (mean±standard deviation) of the total available score in each specialty. P values were derived from independent t-tests.

	Pre-Test (n = 77)	Post-Test (n = 79)	p value
Anaesthesiology & emergency medicine (15 items)	71.9 ± 13.0	78.1 ± 10.6	0.001
General practice (15 items)	71.5 ± 12.9	75.4 ± 10.3	0.041
Paediatrics (15 items)	75.4 ± 10.6	79.4 ± 10.9	0.022
Obstetrics & gynaecology (10 items)	60.0 ± 15.7	68.5 ± 14.6	0.001
Neurology (15 items)	59.8 ± 12.3	67.7 ± 11.8	<0.001
Internal medicine (36 items)	58.0 ± 8.7	64.1 ± 8.2	<0.001
Surgery (15 items)	60.1 ± 11.2	65.9 ± 10.8	0.001
Pharmacology (10 items)	59.9 ± 15.4	65.1 ± 12.9	0.024
Psychiatry (15 items)	61.0 ± 12.0	63.4 ± 11.1	0.194
Urology (4 items)	76.9 ± 20.6	77.8 ± 19.2	0.778

in this study indicates that this formative test provided a valid estimate of student performance levels. In our pre-test sample (i.e., students who had just completed a five-year undergraduate medical curriculum and who self-selected to participate in the time-consuming activity of taking a 150-item examination), the average percent score was as low as 64%. As a percent score of 60% is usually needed to pass the German NLE, this would translate into a 75% pass rate (58 out of 77 students taking the pre-test) in this highly motivated sample. At the end of the practice year, the average percent score was below 70%, and five out of 79 students taking the post-test scored less than 60% of the available points. While it is somewhat reassuring to see that a majority of students would pass the exam used in this study even before entering the practice year, the moderate performance in the post-test is an important finding as the average percent score reported for the spring 2012 NLE in

Germany was 79.4% [http://www.impp.de/IMPP2010/pdf/ErgMedF12.pdf].

At first glance, the increase in performance from the pre- to the post-test might seem surprising as exposure to clinical practice during the final year of medical education would not be expected to increase factual knowledge as assessed in the test due to a lack of constructive alignment between teaching and assessment format [13]. However, it may be hypothesized that students working on the wards encounter a number of opportunities to increase their factual knowledge. Dealing with clinical cases prompts students to build on and expand their knowledge, particularly in the presence of experienced physicians who provide informal teaching during ward rounds. In addition, being involved in patient care is likely to increase student motivation to learn. Vice versa, our finding of students achieving particularly high scores in items related to their specific specialties might reflect a higher motivation to engage in self-directed learning activities regarding their chosen specialty rotation (i.e., higher a priori motivation). Curricular representation of these subjects (i.e., paediatrics, obstetrics and gynaecology) might be partially responsible for this finding. Alternatively, additional teaching in these specialties during the practice year might have been of particularly high quality. However, we did not formally assess teaching quality in the seven different specialties chosen by study participants.

Strengths and limitations

We used specific quality criteria during the selection process for items to be included in the test [14]:

1. selected MCQs were to test important material which was appropriate for the level of training,
2. the stems of the MCQs were mostly case vignettes which contained the majority of information in a focused manner,
3. the five answers were homogeneous in content, length, and grammar.

Even though we used USMLE type items that study participants were not familiar with, this is unlikely to have significantly impacted on our results as this 'origin bias' has been shown to be small when new item formats are presented to advanced undergraduate medical students [15]. Despite careful selection of test items resulting in a balanced examination, item characteristics and internal consistency of the exam were suboptimal. However, similar Cronbach's α values have been reported in a study on different multiple-choice question formats [16], and a recent analysis of data derived from 10 years of postgraduate progress testing in obstetrics and gynaecology yielded even lower values [17]. We cannot pin-point the reason for the relatively low internal consistency of our test and the ones referred to above. At the very least, this appears to be a problem frequently encountered with formative tests of clinical knowledge. It might be hypothesized that the student group self-selecting for participation in our study was relatively homogeneous with respect to performance levels. This might have reduced the variance in test results and thus decreased Cronbach's alpha. More research is needed to determine the impact of the heterogeneity of participating students and/or of the items included in the test on its psychometric properties.

Testing bias (i.e., better performance in the post-test due to students having seen the same items in the pre-test) is unlikely as the pre- and post-tests were taken many months apart and answers had not been distributed after the pre-test. We initially chose a longitudinal design for this study. Power analysis was based on an anticipated response rate of 20%, but the longitudinal sample contained only 16.4% of eligible students. In order to assess the primary endpoint of the study, all students providing data at both time points were included in the analysis, yielding a response rate of over 25%. As a consequence, the test used in this study cannot be referred to as a true 'progress test' as this would have required all students to participate in both tests. The overall small response rate raises concern that selection bias might have impacted on study results. Accordingly, we cannot rule out the possibility that the difference between pre- and post-test results is due to students taking the post-test having genuinely higher performance levels than students taking the pre-test. However, given the overlap between the two cohorts, this is unlikely to completely explain the difference.

Conclusions

A US-based test for medical knowledge used in this study provides a tool for formative assessment of the progress of knowledge for undergraduate medical students in their final practice year. The pre-test may be used by students to guide their individual learning behaviour during the practice year, and the post-test could help them to identify specific areas in which more thorough preparation may be necessary to improve their basic and clinical knowledge.

Acknowledgements

The authors thank all participating panel members in Utrecht and Hamburg and all participating medical students and the administrative team members in Göttingen and Hamburg.

Conflict of interest

The authors declare that they have no competing interests.

References

1. Berkel HJ, Nuy HJ, Geerlings T. The influence of progress tests and block tests on study behaviour. Instruct Sci. 1994;22:317-333. DOI: 10.1007/BF00891784
2. van der Vleuten CP. National, European licensing examinations or none at all? Med Teach. 2009;31(3):189-191.DOI: 10.1080/01421590902741171
3. Coombes L, Ricketts C, Freeman A, Stratford J. Beyond assessment: feedback for individuals and institutions based on the progress test. Med Teach. 2010;32(6):486-490. DOI: 10.3109/0142159X.2010.485652
4. Williams RG, Klamen DL, White CB, Petrusa E, Fincher RM, Whitfield CF, Shatzer JH, McCarty T, Miller BM. Tracking development of clinical reasoning ability across five medical schools using a progress test. Acad Med. 2011;86(9):1148-1154. DOI: 10.1097/ACM.0b013e31822631b3
5. Nouns ZM, Georg W. Progress testing in German speaking countries. Med Teach. 2010;32(6):467-470. DOI: 10.3109/0142159X.2010.485656
6. Vogel D, Gierk B, ten Cate O, Harendza S. Composition of an international medical knowledge test for medical students near graduation. Dundee: AMEE; 2011. Abstract book page 71.
7. Downing SM. Assessment of knowledge with written test formats. In: Norman G, van der Vleuten C, Newble D (Hrsg). International handbook of research in medical education. Dordrecht: Kluwer; 2002. S.647-672. DOI: 10.1007/978-94-010-0462-6_25
8. Schwartz PL, Crooks TJ, Sein KT. Test-retest reliability of multiple true-false questions in preclinical medical subjects. Med Educ. 1986;20(5):399-406. DOI: 10.1111/j.1365-2923.1986.tb01184.x
9. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. 3rd edition ed. Philadelphia: National Board of Medical Examiners; 2001.
10. Le T, Vieregger K. First aid Q & A for the USMLE step 2 CK. 2nd edition ed. New York: McGraw-Hill; 2010.
11. Cohen J. A Power Primer. Psychological Bulletin. 1992;112(1):155-159. DOI: 10.1037/0033-2909.112.1.155
12. Raupach T, Hanneforth N, Anders S, Pukrop T, Th J ten Cate O, Harendza S. Impact of teaching and assessment format on electrocardiogram interpretation skills. Med Educ. 2010;44(7):731-740. DOI: 10.1111/j.1365-2923.2010.03687.x
13. Kern DE, Thomas PA, Howard DM, Bass EB. Curriculum development for medical education - A six-step approach. Baltimore, London: The John Hopkins University Press; 1998.
14. Boland RJ, Lester NA, Williams E. Writing multiple-choice questions. Acad Psychiatry. 2010;34(4):310-316. DOI: 10.1176/appi.ap.34.4.310
15. Muijtjens AM, Schuwirth LW, Cohen-Schotanus J, van der Vleuten CP. Origin bias of test items compromises the validity and fairness of curriculum comparisons. Med Educ. 2007c;41(12):1217-1223.
16. Coderre SP, Harasym P, Mandin H, Fick G. The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. BMC Med Educ. 2004;4:23. DOI: 10.1186/1472-6920-4-23
17. Dijksterhuis MG, Scheele F, Schuwirth LW, Essed GG, Nijhuis JG, Braat DD. Progress testing in postgraduate medical education. Med Teach. 2009;31(10):e464-468. DOI: 10.3109/01421590902849545

Corresponding author:

Tobias Raupach, MD, MME
University Medical Centre Göttingen, Department of Cardiology and Pneumology, 37099 Göttingen, Germany,
Phone: +49 (0)551/39-8922, Fax: +49 (0)551/39-6887
raupach@med.uni-goettingen.de

Please cite as

Raupach T, Vogel D, Schiekirka S, Keijsers C, Ten Cate O, Harendza S. Increase in medical knowledge during the final year of undergraduate medical education in Germany. GMS Z Med Ausbild. 2013;30(3):Doc33. DOI: 10.3205/zma000876, URN: urn:nbn:de:0183-zma0008769

This article is freely available from

http://www.egms.de/en/journals/zma/2013-30/zma000876.shtml

Received: *2012-11-23*
Revised: *2013-03-31*
Accepted: *2013-05-02*
Published: *2013-08-15*

Wissenszuwachs im Praktischen Jahr des Medizinstudiums in Deutschland

Zusammenfassung

Zielsetzung: In Deutschland besteht das letzte Jahr des Studiums der Humanmedizin ("Praktisches Jahr", PJ) aus drei Tertialen von je 16 Wochen, von denen eines in der Inneren Medizin und eines in der Chirurgie abzuleisten ist. Die Fachrichtung des dritten Tertials von 16 Wochen kann von den Studierenden frei gewählt werden. Während des Praktischen Jahres findet keine gezielte Vorbereitung auf den schriftlichen Teil des Staatsexamens statt. Es ist unklar, inwieweit die Studierenden während des Praktischen Jahres auch neue Wissensinhalte erlernen. Ziel dieser Studie war es, zu Beginn und am Ende des Praktischen Jahres Wissensinhalte zu überprüfen.
Methodik: Zehn Absolventen des Medizinstudiums in Deutschland und den Niederlanden trafen eine Auswahl aus 300 vorselektierten Fragen der US-amerikanischen Abschlussprüfung (USMLE), die zehn verschiedenen Fachrichtungen zugeordnet waren. Die ausgewählten 150 Fragen wurden im Rahmen zweier Tests PJ-Studierenden aus Göttingen und Hamburg vorgelegt: einmal zu Beginn (n=77 Studierende) und einmal am Ende des Praktischen Jahres (n=79).
Ergebnisse: Die interne Konsistenz der beiden Tests (Cronbach's α) betrug 0,75 (Prätest) bzw. 0,68 (Posttest). Der Anteil richtig beantworteter Fragen betrug im Prätest 63,9±6,9 und im Posttest 69,4±5,7 ($p<0,001$; Effektstärke als Cohen's d: 0,87). Individuelle Studierende schnitten bei denjenigen Fragen besonders gut ab, die sich auf Inhalte ihres Wahlfachs bezogen.
Schlussfolgerung: Der in dieser Studie verwendete Wissenstest eignet sich als externes Instrument zur Messung des Wissenszuwachses von Studierenden im Praktischen Jahr. Zudem kann der Prätest genutzt werden, um Studierende bei der Planung ihres Lernverhaltens während des Praktischen Jahres zu unterstützen.

Schlüsselwörter: Medizinstudium, Klinik, Wissen

Tobias Raupach[1]
Daniela Vogel[2]
Sarah Schiekirka[1,3]
Carolina Keijsers[4]
Olle Ten Cate[5]
Sigrid Harendza[2]

1 Universitätsmedizin Göttingen, Abteilung Kardiologie & Pneumologie, Göttingen, Deutschland
2 Universitätsklinikum Hamburg-Eppendorf, III. Medizinische Klinik, Hamburg, Deutschland
3 Universitätsmedizin Göttingen, Studiendekanat, Göttingen, Deutschland
4 Universitätsmedizin Utrecht, Abteilung für Geriatrische Medizin, Utrecht, Niederlande
5 Universitätsmedizin Utrecht, Zentrum für Ausbildungsforschung und -entwicklung, Utrecht, Niederlande

Einleitung

Deutsche Studierende der Humanmedizin rotieren im letzten Jahr des Studiums (Praktisches Jahr, PJ) durch verschiedene klinische Bereiche. Während alle Studierenden zwei der drei Tertiale für jeweils 16 Wochen in den Fächern Innere Medizin und Chirurgie ableisten müssen, kann sich jeder Studierende für ein Wahlfach für die übrigen 16 Wochen des PJ entscheiden. Bis zum Jahr 2012 fand nach dem PJ das aus 320 Multiple Choice (MC)-Fragen bestehende Zweite Staatsexamen statt, in dem sowohl Grundlagen als auch klinische Wissensinhalte abgefragt wurden. Eine spezifische Vorbereitung auf dieses Examen fand in aller Regel während des PJs nicht statt. Folglich waren die Studierenden gezwungen, während dieses letzten Jahres einerseits klinische Erfahrung zu sammeln und andererseits den für das Bestehen des Staatsexamens erforderlichen theoretischen Wissensstand zu erreichen. Da während des PJs keine Pflichtprüfungen stattfinden, war den Studierenden eine objektive Überprüfung ihres Lernerfolgs bislang nicht möglich.
Mit dem Einsatz von Progress Tests im Medizinstudium sollen Lernprozesse gefördert werden, die stärker auf das Verständnis und weniger auf die reine Reproduktion von Inhalten abzielen [1]. Zur Messung des studentischen Lernerfolgs hinsichtlich medizinischen Faktenwissens wurden in verschiedenen Ländern solche Progress Tests entwickelt [2], [3], [4]. Einige Studierende nutzen das durch Progress Tests generierte Feedback zur Erstellung individueller Lernpläne [5]. Entsprechend könnten formative Wissenstests zu Beginn und am Ende des PJs sich als hilfreiche Feedback-Instrumente für Studierende er-

weisen. In dieser Studie wurde zu diesem Zweck ein aus 150 Items bestehender MC-Test verwendet. Dieser Test war ursprünglich dazu entwickelt worden, den Leistungsstand von Studierenden unterschiedlicher Europäischer Länder miteinander zu vergleichen [6]. Im Rahmen der vorliegenden Arbeit wurde er an zwei deutschen Universitäten eingesetzt. Das primäre Ziel der Studie war die Messung des Wissenszuwachses während des PJ. Wir stellten die Hypothese auf, dass die studentischen Leistungen am Ende des PJs signifikant höher liegen als am Anfang, da die Exposition gegenüber dem klinischen Alltag auch die Aneignung von Faktenwissen fördern könnte. Es wurde erwartet, dass dieser Effekt besonders für diejenigen Wissensinhalte nachzuweisen sein würde, die sich auf das jeweilige Wahlfach eines Studierenden bezogen.

Methodik

Verwendung von MC-Fragen

Mit Hilfe von MC-Fragen kann Faktenwissen valide geprüft werden [7]. MC-Prüfungen können zudem eine hohe Reliabilität erreichen, was eine wesentliche Voraussetzung dafür ist, dass eine Prüfung zwischen leistungsstarken und leistungsschwachen Studierenden unterscheiden kann [8]. Mit MC-Fragen kann nicht nur die Fähigkeit zur Reproduktion von Wissensinhalten geprüft werden, sondern es ist mit diesem Format auch möglich, komplexere kognitive Prozesse wie z.B. die Interpretation oder Anwendung von Wissen abzubilden [9]. Für MC-Fragen wurden unterschiedliche Formate entwickelt, die sich insbesondere hinsichtlich der Anzahl der Antwortoptionen unterscheiden. Sowohl im deutschen Staatsexamen als auch im amerikanischen Examen (United States Medical Licensing Examination, USMLE) werden typischerweise Fragen mit einer richtigen Antwort und vier Distraktoren verwendet.

Auswahl der Prüfungsfragen

Der aus 150 Fragen bestehende Wissenstest wurde aus 1000 frei verfügbaren USMLE-äquivalenten Items zusammengestellt [10]. In einem ersten Schritt waren 300 Fragen nach den folgenden Kriterien vorselektiert und adaptiert worden:

1. Alle Fragen mussten zu einer der folgenden Fachrichtungen gehören: Allgemeinmedizin, Anästhesiologie und Notfallmedizin, Innere Medizin, Chirurgie, Urologie, Gynäkologie, Pädiatrie, Neurologie, Psychiatrie, klinische Pharmakologie. Kleinere Fächer wie beispielsweise Hals-Nasen-Ohrenheilkunde und Dermatologie wurden nicht berücksichtigt.
2. Alle Fragen basierten auf klinischen Fällen.
3. Erkrankungen, die spezifisch im amerikanischen Sprachraum auftreten (z.B. „Rocky Mountain spotted fever“), wurden nicht eingeschlossen.
4. Fragen mit Bildbeilage (Röntgenbilder, EKG-Registrierungen) konnten aus Copyright-Gründen nicht berücksichtigt werden.
5. Alle Antwortoptionen einer Frage mussten zur gleichen Kategorie gehören (z.B. diagnostische, therapeutische oder andere Maßnahmen). Fragen, die beispielsweise zwei diagnostische und drei therapeutische Antwortoptionen beinhalteten, wurden entweder nicht eingeschlossen oder entsprechend modifiziert.
6. Wenn mehrere Fragen dasselbe Thema abdeckten, wurde jeweils nur eine ausgewählt.

Im Februar 2011 wurden die 300 vorselektierten Fragen von jeweils fünf Absolventen des Medizinstudiums an den Universitäten in Hamburg und in Utrecht (Niederlande) daraufhin begutachtet, ob sie in mehr als einem europäischen Land einsetzbar sind. Hierzu wählte jede/r der zehn Gutachter/innen diejenigen 150 Fragen aus, die ihrer/seiner Meinung nach am ehesten dem Wissensstand von Studienabsolventen ihres Herkunftslandes entsprachen. Im Interesse einer hohen Inhaltsvalidität war dabei die Fragenzahl pro Fachrichtung vorgegeben. Der endgültige Test bestand aus den 150 Fragen, die von mindestens zwei der fünf Gutachter/innen aus jedem der beiden Länder gewählt worden waren. Jeweils ein/e Gutachter/in aus Hamburg und Utrecht nahm dann noch eine Bewertung der Fragen hinsichtlich ihrer Schwierigkeit und ihrer inhaltlichen Domäne (Grundlagen bzw. klinische Wissensinhalte) zur Brauchbarkeitstestung des endgültigen Tests vor. Dies resultierte in einem adäquat balancierten Test mit einer Intraklassenkorrelation von 0,85 bzw. 0,71. Im Rahmen einer Pilotierung an 56 Studierenden aus Deutschland und den Niederlanden wurde die interne Konsistenz des Tests mit einem Cronbach's α von 0,79 ermittelt [6].

Durchführung der Prä- und Posttests

Im Rahmen der vorliegenden Studie wurde eine deutsche Version der 150 Fragen im Rahmen zweier formativer Prüfungen eingesetzt. Zum Prätest wurden 286 PJ-Studierende (164 in Göttingen, 122 in Hamburg) eingeladen, die im April 2011 am Anfang des Praktischen Jahres standen. Die Studierenden erhielten eine E-Mail, in der Sinn und Zweck der Studie dargestellt wurden. Entsprechend einer Vereinbarung mit den Lehrkrankenhäusern wurden die Studien-Teilnehmer/innen für die Zeit des Prätests von ihren klinischen Aufgaben entbunden.
Alle Studierenden, die im April 2011 zum Prätest eingeladen worden waren, erhielten am Ende ihres Praktischen Jahres (d.h. vor Beginn ihrer Vorbereitung auf das Staatsexamen im Frühjahr 2012) eine Einladung zur Teilnahme am Posttest, der dieselben Fragen enthielt wie der Prätest. Die Fragen und Antworten wurden nicht an die Studierenden ausgegeben und alle Papiere wurden nach dem Test eingesammelt. Allerdings wurden alle Studienteilnehmer/innen per E-Mail über ihre individuellen Ergebnisse informiert. Studienteilnehmer/innen wurden um Angaben zu sozioökonomischen Daten, zu einem eventuellen Auslandsstudium und zu ihren PJ-Wahlfä-

chern gebeten. Die Teilnahme an der Studie war freiwillig. Die Studie wurde von der Ethikkommission der Ärztekammer Hamburg genehmigt.

Datensammlung und statistische Analyse

Die studienbezogenen Fragebögen und die Prä- und Posttests wurden manuell in eine SPSS-Datenbank übertragen (SPSS 19.0; SPSS Inc., Chicago, Illinois, USA). Unterschiede zwischen den beiden Studierendengruppen, die jeweils am Prä- und Posttest teilgenommen hatten, wurden mittels χ^2-Tests (dichotome Variablen) und t-Tests (kontinuierliche Variablen) auf Signifikanz untersucht. Effektstärken wurden als Cohen's d berechnet; hierbei entsprechen Werte von 0,2 kleinen und Werte von 0,8 großen Effekten [11]. Sowohl für den Prä- als auch für den Posttest wurden die gängigen Testgütekriterien (Item-Schwierigkeit, Trennschärfe und Cronbach's α als Maß für die interne Konsistenz) bestimmt. Zur Detektion eines Leistungsunterschieds von drei Prozentpunkten zwischen zwei Gruppen (z.B. 68% versus 65% bei angenommenen Standardabweichungen von 6,5% in beiden Gruppen) auf einem Signifikanzniveau von 5% und einer Power von 80% mussten sowohl am Prä- als auch am Posttest mindestens 58 Studierende teilnehmen (äquivalent zu einer longitudinalen Teilnahmequote von 20%). Da ein Unterschied um drei Prozentpunkte im deutschen Benotungssystem etwa eine Drittel-Note ausmacht, wurde diese Differenz als bedeutsam erachtet. Die Ergebnisse sind als Mittelwerte±Standardabweichungen oder als prozentuale Anteile (n) angegeben. Das Signifikanzniveau wurde auf $p<0,05$ festgelegt.

Ergebnisse

Teilnahmequote und Teilnehmer-Charakteristika

Am Prätest nahmen 77 Studierende teil, am Posttest 79 Studierende (Teilnahmequoten jeweils 26,9% und 27,6%). Der Anteil weiblicher Studierender im Prä- und Posttest betrug 66,2% und 73,4%. Die Teilnahmequoten waren in Hamburg höher und schwankten an beiden Studienorten zwischen dem Prä- und dem Posttest (Hamburg: 45/122 (36,9%) versus 58/122 (47,5%); Göttingen: 32/164 (19,5%) versus 21/164 (12,8%)). Insgesamt 47 Studierende nahmen sowohl am Prä- als auch am Posttest teil. Die Charakteristika der Studienteilnehmer/innen sind in Tabelle 1 dargestellt. Wie erwartet waren die Studierenden im Posttest signifikant älter als diejenigen im Prätest. Zwischen den beiden Kohorten ergaben sich keine signifikanten Unterschiede hinsichtlich des Geschlechts, der Muttersprache, einer vorhergehenden Berufsausbildung, der Ableistung von Teilen des PJs im Ausland und des Wahlfachs.

Itemanalyse

Das Cronbach's α des Prä- und Posttest betrug 0,75 bzw. 0,68. Die Schwierigkeit einzelner Fragen des Prätests lag zwischen 0,03 und 1,00 (Mittelwert 0,64) und für den Posttest zwischen 0,04 bis 1,00 (Mittelwert 0,69). Der Anteil der Fragen mit einer Schwierigkeit zwischen 0,4 und 0,8 betrug im Prätest 56,7% (n=85) und im Posttest 50,7% (n=76). Die korrigierten Trennschärfen einzelner Items lagen im Prätest zwischen -0,20 und 0,39 (Mittelwert 0,13) und im Posttest zwischen -0,32 und 0,45 (Mittelwert 0,10). Der Anteil der Fragen mit einer positiven Trennschärfe betrug im Prätest 85,3% (n=128) und im Posttest 70,7% (n=106).

Studentische Leistungen

Von den Studierenden wurden im Prätest ein mittlerer Prozentwert richtig beantworteter Fragen von 63,9±6,9 und im Posttest von 69,4±5,7% erreicht ($T(154)=-5,376$; $p<0,001$; t-Test für unabhängige Stichproben). Die Effektstärke dieses Unterschieds, berechnet als Cohen's d, betrug 0,87, was einem großen Effekt entspricht. Eine Berechnung für verbundene Stichproben, in der nur die 47 Studierenden berücksichtigt wurden, die an beiden Tests teilgenommen hatten, lieferte ein vergleichbares Ergebnis: 64,6±6,7% im Prä- und 69,6±5,3 im Posttest ($T(46)=-7,299$; $p<0,001$). Aus der nach Fachrichtungen aufgeschlüsselten Analyse der Testleistungen (siehe Tabelle 2) geht hervor, dass Studierende mit dem Wahlfach „Anästhesiologie und Notfallmedizin" im Posttest in den Fragen mit Bezug zu diesem Fachgebiet nicht besser abschnitten als Studierende mit einem anderen Wahlfach (11,7±1,4 bzw. 11,7±1,6 von 15 erreichbaren Punkten; $p=0,985$); ein ähnliches Muster ergab sich für Studierende mit dem Wahlfach Neurologie (9,9±2,0 bzw. 10,2±1,8 von 15 Punkten; $p=0,639$). Im Gegensatz dazu erreichten Studierende mit dem Wahlfach Pädiatrie im Posttest in den fachbezogenen Fragen höhere Punktzahlen als Studierende mit einem anderen Wahlfach (13,6±1,3 vs. 11,6±1,5 von 15 Punkten; $p<0,001$; Cohen's $d=1,40$) und das gleiche galt auch für Studierende mit dem Wahlfach Gynäkologie (7,8±1,3 vs. 6,7±1,4 von 10 Punkten; $p=0,017$; Cohen's $d=0,81$). Es fiel auf, dass fünf der 14 Studierenden mit dem Wahlfach Pädiatrie alle Fragen aus diesem Fachgebiet richtig beantworteten – dies gelang hingegen nur einem der 65 Studierenden mit einem anderen Wahlfach.

Diskussion

Mit Hilfe eines aus 150 Fragen bestehenden und an das Format des USMLE angelehnten formativen Prä-Post-Tests für PJ-Studierende in Deutschland wurde in dieser Studie eine signifikante Zunahme des Wissensstandes nach dem PJ nachgewiesen. Der Lernzuwachs war auf denjenigen Gebieten besonders stark ausgeprägt, mit denen die Studierenden sich im Rahmen ihrer Wahlfächer

Tabelle 1: Charakteristika der Teilnehmenden. Die p-Werte beziehen sich auf einen t-Test für unabhängige Stichproben (Alter) bzw. auf χ^2-Tests (alle übrigen Variablen).

	Prätest (n = 77)	Posttest (n = 79)	p-Wert	Prä- and Post-Test (n = 47)
Alter, Mittelwert ± Standardabweichung	26,5 ± 2,1	27,3 ± 3,0	**0,048**	26,4 ± 2,3
Anteil weiblicher Studierender, % (n)	66,2 (51)	73,4 (58)	0,328	74,5 (35)
Anteil der Studierenden mit Deutsch als Muttersprache, % (n)	84,4 (65)	88,6 (70)	0,443	91,5 (43)
Anteil der Studierenden mit abgeschlossener Berufsausbildung, % (n)	16,4 (10)	17,7 (14)	0,836	19,1 (9)
Anteil der Studierenden, die teilweise im Ausland studiert haben, % (n)	21,3 (13)	30,4 (24)	0,228	21,3 (10)
Wahlfach-Verteilung				
Anästhesiologie & Notfallmedizin, % (n)	19,7 (12)	17,7 (14)	0,620	14,9 (7)
Pädiatrie, % (n)	19,7 (12)	17,7 (14)		19,1 (9)
Gynäkologie, % (n)	16,4 (10)	13,9 (11)		14,9 (7)
Neurologie, % (n)	11,5 (7)	11,4 (9)		14,9 (7)
Hals-Nasen-Ohrenheilkunde, % (n)	3,3 (2)	8,9 (7)		4,3 (2)
Urologie, % (n)	0,0 (0)	5,1 (4)		0 (0)
Radiologie, % (n)	4,9 (3)	5,1 (4)		6,4 (3)
Sonstige, % (n)	24,6 (15)	20,3 (16)		25,5 (12)

Tabelle 2: Studentische Leistungen im Prä- und Posttest, aufgeschlüsselt nach Fächern. Die Leistung ist in Prozent (Mittelwert±Standardabweichung) der erreichbaren Punkte angegeben. Die p-Werte beziehen sich auf t-Tests für unabhängige Stichproben.

	Prätest (n = 77)	Posttest (n = 79)	p-Wert
Anästhesiologie & Notfallmedizin (15 Fragen)	71,9 ± 13,0	78,1 ± 10,6	0,001
Allgemeinmedizin (15 Fragen)	71,5 ± 12,9	75,4 ± 10,3	0,041
Pädiatrie (15 Fragen)	75,4 ± 10,6	79,4 ± 10,9	0,022
Gynäkologie (10 Fragen)	60,0 ± 15,7	68,5 ± 14,6	0,001
Neurologie (15 Fragen)	59,8 ± 12,3	67,7 ± 11,8	<0,001
Innere Medizin (36 Fragen)	58,0 ± 8,7	64,1 ± 8,2	<0,001
Chirurgie (15 Fragen)	60,1 ± 11,2	65,9 ± 10,8	0,001
Pharmakologie (10 Fragen)	59,9 ± 15,4	65,1 ± 12,9	0,024
Psychiatrie (15 Fragen)	61,0 ± 12,0	63,4 ± 11,1	0,194
Urologie (4 Fragen)	76,9 ± 20,6	77,8 ± 19,2	0,778

intensiver auseinandergesetzt hatten. Eine Abschätzung des eigenen Wissensstandes im Rahmen eines formativen Tests kann Studierende bei der Erstellung ihres Lernplans für das PJ unterstützen. Ein solcher Test war bisher nicht verfügbar. Das hier vorgestellte, neu entwickelte Instrument zur formativen Prüfung von PJ-Studierenden könnte diese Lücke schließen. Einige Teilnehmer/innen dieser Studie erklärten im Rahmen eines informellen Feedbacks, dass sie es als hilfreich empfunden hätten, ihr eigenes Wissen unter simulierten Prüfungsbedingungen testen zu können und dass sie die Ergebnisse des Prätests in der Tat zur Planung ihres Lernverhaltens im PJ nutzten.

Leistungszuwachs im PJ

Für die vergleichsweise niedrigen Leistungen im Posttest sind verschiedene Erklärungen denkbar. Der formative Charakter der Prä- und Posttests ging möglicherweise mit einem geringeren Leistungsanreiz einher als wenn beide Prüfungen einen summativen Ansatz verfolgt hätten, was folglich zu einer möglicherweise falsch-niedrig

gemessenen Erfassung des tatsächlichen Leistungsniveaus der Studierenden geführt haben könnte [12]. Andererseits ließe sich auch hypothetisieren, dass Studienteilnehmer/innen besonders an ihrem eigenen Leistungsstand interessiert waren und sich daher in beiden Tests besonders angestrengt haben, um möglichst alle Fragen richtig zu beantworten. Möglicherweise waren die Studierenden jedoch mit der Formulierung der USMLE-artigen Fragen noch nicht hinreichend vertraut.

Selbst in Anbetracht dieser möglichen Limitationen deutet der von uns beobachtete Anstieg der Leistung vom Präzum Posttest darauf hin, dass der genutzte formative Test eine valide Einschätzung der studentischen Leistungsfähigkeit ermöglicht. In unserer Prätest-Kohorte (d.h. Studierende, die gerade ein fünfjähriges medizinisches Curriculum hinter sich hatten und freiwillig an der zeitaufwendigen Aktivität einer 150-Fragen Prüfung teilnahmen) wurden im Mittel nur 64% der Punkte erreicht. Unter der Annahme einer Bestehensgrenze von 60% ergibt sich aus unseren Daten eine Bestehensquote von 75% (58 der 77 Prätest-Teilnehmer/innen) in dieser hochmotivierten Subgruppe. Am Ende des Praktischen Jahrs wurden weiterhin im Mittel weniger als 70% der Punkte erreicht, und fünf der 79 Posttest-Teilnehmer/innen erreichten weniger als 60% der möglichen Punkte. Einerseits ist positiv anzumerken, dass die Mehrheit der Studierenden den in dieser Studie genutzten Test bereits zu Beginn des PJs bestanden hätte. Andererseits ist die moderate Leistung im Posttest ein wichtiger Befund, da im Frühjahr 2012 von deutschen Studierenden im Zweiten Staatsexamen Medizin 79,4% der Punkte erreicht wurden [http://www.impp.de/IMPP2010/pdf/ErgMedF12.pdf].

Auf den ersten Blick mag die Leistungszunahme zwischen dem Prä- und dem Posttest erstaunen, da nicht zwingend erwartet werden kann, dass der für das PJ typische Einsatz in einem klinischen Umfeld aufgrund von mangelnder Kongruenz zwischen Lehr- und Prüfungsform auch mit einer Zunahme von Faktenwissen einhergeht [13]. Allerdings ist anzunehmen, dass Studierende während ihrer klinischen Ausbildung zahlreiche Gelegenheiten zum Ausbau ihres Faktenwissens wahrnehmen. So gibt die Befassung mit klinischen Fällen nicht selten Anlass zur Auffrischung von bereits Gelerntem und zur Aufnahme neuer Inhalte, insbesondere, wenn erfahrene Kliniker im Rahmen der Visite informelle Lehre anbieten. Schließlich dürfte die Beteiligung von Studierenden an der Krankenversorgung die Lernmotivation per se erhöhen. Umgekehrt lassen die hier beobachteten besonders guten Leistungen von Studierenden bei Fragen mit Bezug zu ihrem individuellen Wahlfach darauf schließen, dass Studierende primär eine höhere Lernmotivation für das von ihnen selbst gewählte Fach aufweisen. Unsere diesbezüglichen Ergebnisse sind vor dem Hintergrund der curricularen Verankerung der beiden Fächer Pädiatrie und Gynäkologie zu interpretieren. Zudem war die PJ-begleitende Lehre in diesen Fächern möglicherweise besonders gut. Die Lehrqualität in den sieben von den Teilnehmern absolvierten Wahlfächern wurde in der vorliegenden Studie allerdings nicht untersucht.

Stärken und Schwächen

Für das oben beschriebene Fragen-Auswahlverfahren wurden spezifische Qualitätskriterien definiert [14]:

1. die genutzten Fragen mussten wichtige Inhalte abbilden, die dem Ausbildungsstand der PJ-Studierenden angemessen waren,
2. die Fragenstämme bestanden meist aus Fallvignetten, in denen die wesentlichen Informationen fokussiert dargestellt waren,
3. die fünf Antwortoptionen waren bezüglich ihrer Inhalte, Länge und Grammatik homogen.

Dass die Studienteilnehmer/innen noch nicht eng mit dem an das USMLE angelehnte Fragenformat vertraut waren, dürfte auf die Ergebnisse eher von geringem Einfluss gewesen sein, da gezeigt wurde, dass der aus der Präsentation neuer Fragenformate resultierende „Herkunftsbias“ bei fortgeschrittenen Studierenden keine große Rolle spielt [15]. Trotz der sorgfältigen Fragenauswahl und der Konstruktion einer balancierten Prüfung fielen die Itemkennwerte und die interne Konsistenz beider Tests suboptimal aus. Allerdings ergaben sich in einer Untersuchung verschiedener MC-Fragenformate ähnliche Werte für Cronbach's α [16] und eine noch geringere interne Konsistenz fand sich in einer kürzlich publizierten Arbeit zu 10-jährigem postgraduiertem Progress Testing in der Gynäkologie [17]. Die Ursache für die relativ geringe interne Konsistenz unseres und der hier zitierten Tests bleibt letztlich unklar. Offenbar tritt dieses Problem im Kontext klinischer Wissenstests jedoch nicht selten auf. Es ist denkbar, dass Studierende, die sich freiwillig zur Teilnahme an einer Studie melden, insgesamt ein relativ homogenes Leistungsniveau aufweisen. Die daraus resultierende geringere Varianz der Testergebnisse geht mit einem geringeren Cronbach's α einher. Künftige Forschungsprojekte sollten den Einfluss der Heterogenität der Studienteilnehmer und/oder der genutzten Prüfungsfragen auf die psychometrischen Eigenschaften eines Prüfungsinstruments näher untersuchen.

Ein eventueller „testing bias“ (d.h. Studierende erzielen bessere Leistungen im Posttest, weil sie sich an die Fragen des Prätests erinnern) dürfte in unserer Studie eher gering ausgefallen sein, da mehrere Monate zwischen den beiden Tests lagen und weder die Fragen noch die richtigen Antworten veröffentlicht wurden. Zur Bearbeitung des primären Studienziels wählten wir ein longitudinales Design. In der Power-Analyse wurde eine Teilnahmequote von 20% angenommen, aber die Studierenden-Stichprobe mit kompletten longitudinalen Daten fiel mit 16,4% der eingeladenen Studierenden etwas geringer aus. Um die Annahme der Power-Analyse nicht zu verletzen, wurden Daten von allen Studierenden ausgewertet, die zu mindestens einem Zeitpunkt an dem Test teilgenommen hatten (Teilnahmequote 25%). Somit wird in dieser Arbeit kein „Progress Test“ im engeren Sinne vorgestellt, da hierzu alle Studierenden an beiden Zeitpunkten am Test hätten teilnehmen müssen. Die insgesamt geringe Teilnahmequote legt nahe, dass die Ergebnisse

durch einen Selektionsbias beeinflusst worden sind. Es ist nicht auszuschließen, dass der beobachtete Unterschied zwischen den Leistungen im Prä- und Posttest auch dadurch zustande kam, dass Studierende, die am Posttest teilnahmen, ein höheres Leistungsniveau aufwiesen als Studierende, die am Prätest teilnahmen. Angesichts der großen Überschneidungen zwischen beiden Gruppen ist der Unterschied durch diesen Effekt aber wahrscheinlich nicht komplett zu erklären.

Schlussfolgerung

Der in dieser Studie genutzte und an das amerikanische Examen angelehnte Wissenstest eignet sich zur formativen Erfassung des studentischen Wissenserwerbs im Praktischen Jahr. Der Prätest könnte die Studierenden bei der Planung ihres Lernverhaltens während des Praktischen Jahrs unterstützen und der Posttest könnte der Identifikation von Wissenslücken dienen, die während der Vorbereitung auf das Staatsexamen noch gefüllt werden müssen.

Danksagung

Die Autoren danken allen Mitgliedern des Fragenauswahlkommittees in Utrecht und Hamburg, den teilnehmenden Studierenden und den Organisations-Teams in Göttingen und Hamburg.

Interessenkonflikt

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

Literatur

1. Berkel HJ, Nuy HJ, Geerlings T. The influence of progress tests and block tests on study behaviour. Instruct Sci. 1994;22:317-333. DOI: 10.1007/BF00891784
2. van der Vleuten CP. National, European licensing examinations or none at all? Med Teach. 2009;31(3):189-191.DOI: 10.1080/01421590902741171
3. Coombes L, Ricketts C, Freeman A, Stratford J. Beyond assessment: feedback for individuals and institutions based on the progress test. Med Teach. 2010;32(6):486-490. DOI: 10.3109/0142159X.2010.485652
4. Williams RG, Klamen DL, White CB, Petrusa E, Fincher RM, Whitfield CF, Shatzer JH, McCarty T, Miller BM. Tracking development of clinical reasoning ability across five medical schools using a progress test. Acad Med. 2011;86(9):1148-1154. DOI: 10.1097/ACM.0b013e31822631b3
5. Nouns ZM, Georg W. Progress testing in German speaking countries. Med Teach. 2010;32(6):467-470. DOI: 10.3109/0142159X.2010.485656
6. Vogel D, Gierk B, ten Cate O, Harendza S. Composition of an international medical knowledge test for medical students near graduation. Dundee: AMEE; 2011. Abstract book page 71.
7. Downing SM. Assessment of knowledge with written test formats. In: Norman G, van der Vleuten C, Newble D (Hrsg). International handbook of research in medical education. Dordrecht: Kluwer; 2002. S.647-672. DOI: 10.1007/978-94-010-0462-6_25
8. Schwartz PL, Crooks TJ, Sein KT. Test-retest reliability of multiple true-false questions in preclinical medical subjects. Med Educ. 1986;20(5):399-406. DOI: 10.1111/j.1365-2923.1986.tb01184.x
9. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. 3rd edition ed. Philadelphia: National Board of Medical Examiners; 2001.
10. Le T, Vieregger K. First aid Q & A for the USMLE step 2 CK. 2nd edition ed. New York: McGraw-Hill; 2010.
11. Cohen J. A Power Primer. Psychological Bulletin. 1992;112(1):155-159. DOI: 10.1037/0033-2909.112.1.155
12. Raupach T, Hanneforth N, Anders S, Pukrop T, Th J ten Cate O, Harendza S. Impact of teaching and assessment format on electrocardiogram interpretation skills. Med Educ. 2010;44(7):731-740. DOI: 10.1111/j.1365-2923.2010.03687.x
13. Kern DE, Thomas PA, Howard DM, Bass EB. Curriculum development for medical education - A six-step approach. Baltimore, London: The John Hopkins University Press; 1998.
14. Boland RJ, Lester NA, Williams E. Writing multiple-choice questions. Acad Psychiatry. 2010;34(4):310-316. DOI: 10.1176/appi.ap.34.4.310
15. Muijtjens AM, Schuwirth LW, Cohen-Schotanus J, van der Vleuten CP. Origin bias of test items compromises the validity and fairness of curriculum comparisons. Med Educ. 2007c;41(12):1217-1223.
16. Coderre SP, Harasym P, Mandin H, Fick G. The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. BMC Med Educ. 2004;4:23. DOI: 10.1186/1472-6920-4-23
17. Dijksterhuis MG, Scheele F, Schuwirth LW, Essed GG, Nijhuis JG, Braat DD. Progress testing in postgraduate medical education. Med Teach. 2009;31(10):e464-468. DOI: 10.3109/01421590902849545

Korrespondenzadresse:
Tobias Raupach, MD, MME
Universitätsmedizin Göttingen, Abteilung Kardiologie & Pneumologie, D-37099 Göttingen, Deutschland, Tel.: +49 (0)551/39-8922, Fax: +49 (0)551/39-6887
raupach@med.uni-goettingen.de

Bitte zitieren als
Raupach T, Vogel D, Schiekirka S, Keijsers C, Ten Cate O, Harendza S. Increase in medical knowledge during the final year of undergraduate medical education in Germany. GMS Z Med Ausbild. 2013;30(3):Doc33. DOI: 10.3205/zma000876, URN: urn:nbn:de:0183-zma0008769

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/zma/2013-30/zma000876.shtml

Eingereicht: *23.11.2012*
Überarbeitet: *31.03.2013*
Angenommen: *02.05.2013*
Veröffentlicht: *15.08.2013*