Do they know too little? An inter-institutional study on the anatomical knowledge of upper-year medical students based on multiple choice questions of a progress test

Irene Brunk*, Stefan Schauber[#], Waltraud Georg[+]
*Institut für Integrative Neuroanatomie, Charité-Universitätsmedizin Berlin
[#] Faculty of Educational Sciences/ Faculty of Medicine, University of Oslo
[+] Helios Kliniken GmbH

Abstract

The depth of medical students' knowledge of human anatomy is often controversially discussed. In particular, members of surgical disciplines raise concerns regarding deficits in the factual anatomical and topographical knowledge of upper-year students. The question often raised is whether or not medical students have sufficient anatomical and topographical knowledge when they graduate from medical school. Indeed, this question is highly relevant for curricular planners. Therefore, we have addressed it by evaluating the performance of students in the 5th and 6th years of their studies on anatomical multiple choice questions from the Berlin Progress Test Medicine performed at 10 German university medical schools. Results were compared to a reference based on a standard setting (modified Angoff-procedure). The reference was established independently by 5 panels of anatomists at different universities across Germany. As the ratings were independent of university affiliation, teaching-experience or training of the anatomists, an overall cutting score could be calculated which corresponded to 60.4% correct answers for the question set used in this study.

In the progress test, on average only 29.9% of the students' answers were correct, reflecting that the performance was significantly below the expected standard. On the basis of the test results it remained unclear, whether acquisition or retention of anatomical information was insufficient. Further evaluation by item characteristics revealed that the students had major difficulty in applying their theoretical knowledge to practical problems in

the context of a clinical setting. Thus, our results reveal deficits in the anatomical knowledge of medical students in their final years. Therefore medical curricula should not only focus on enhancing the acquisition and retention of core anatomical knowledge but aim at improving skills applying this in a clinical setting.

Keywords anatomy, undergraduate medical education, knowledge, outcome, surgery, progress testing, standard setting

Introduction

Anatomy is a central component of preclinical medical education and serves as an essential basis for the understanding of the human body, enabling doctors to perform proper physical examination, derive structural diagnosis and apply therapeutic procedures to patients (Mylopoulos and Woods, 2014; Rikers et al., 2005b; Woods, 2007; Woods et al., 2007a). Nevertheless, the anatomical knowledge of medical students has been controversially discussed in recent years (Bergman et al., 2008; Bergman et al., 2011; Prince et al., 2005). In particular, members and associations of surgical disciplines pointed toward deficits in the topographical knowledge of upper-year medical students and young doctors (Chirurgie, 2009; Waterston and Stewart, 2005).

Clinicians from different specialties have repeatedly expressed their impression that anatomical knowledge of medical students is below a minimum level and may even endanger patient safety (Waterston and Stewart, 2005). Some authors propose links between changes in anatomy teaching, the perceived decline in anatomy knowledge of students and young physicians and the increase of reported medical malpractice (Older, 2004; Turney, 2007). Deficits could be attributed to various factors: (1) A major factor is the way how teaching anatomy has changed over the last decades: While changes such as the vertical integration of subjects into the curriculum and an interdisciplinary approach are positively seen and believed to promote retention of knowledge (Bergman et al., 2011), the broad introduction of problem based learning (PBL) raised concerns regarding the acquisition of anatomical/basic science knowledge (Bergman et al., 2014; Cahill et al., 2000; Williams and Lau, 2004) . However, a comparison of students from non-PBL *versus* PBL-curricula revealed no difference in their anatomical knowledge (Prince et al., 2003). In fact, students perceived their knowledge as deficient independent of the type of curriculum they were in (Prince et al., 2003). (2) Some new curriculum dropped dissection classes from their program, despite the fact that the use of human cadavers for teaching anatomy has been found to have a positive impact on the acquisition of topographical as well as general anatomical knowledge (Biasutto et al., 2006; Saltarelli et al., 2014; Winkelmann, 2007). (3) Another factor might be the fact, that anatomy is increasingly taught by non-medically trained staff and the student –staff ratios have severely increased (Pryde and Black, 2005). There is, however, currently no published study investigating whether the field of qualification of teachers has any impact on the anatomy knowledge of students. Intriguingly, complaints about the declining anatomical knowledge of students and graduates were already raised more than forty years ago (Sinclair, 1975). Thus, concerns expressed in the recent years may be independent of the curricular changes occurred in the last couple of decades.

Nevertheless, this ongoing discussion on the appropriateness of graduates' familiarity with anatomy flags up a potential discrepancy between the importance of anatomical knowledge for clinical practice (Rikers et al., 2005a; Woods et al., 2007a, b; Woods et al., 2006) and the outcome of current medical education (Older, 2004). Therefore, it is important to know if the *perceived* deficits can indeed be verified. In the context of the worldwide movement for reforming medical curricula taking place in the last couple of decades, this question has particularly relevance: curricular organizers require information about the acquisition of subject-specific content in different curricula in order to counteract potential deficits. Such information can then form the basis to integrate the various medical disciplines, including anatomy, in such a way that students are best prepared for their future clinical practice. However, before these practical issues can be addressed, obtaining an assessment of students' proficiency is imperative.

Anatomical knowledge, as other medical subjects, is typically measured by various assessment tools such as multiple choice exams, oral exams or objective structured practical examinations (Schoeman and Chandratilake, 2012). One established way to assess factual knowledge and its clinical application are multiple choice items (Wass et al., 2001). Compared to other procedures, one clear benefit of multiple choice items is that it is easier to administer them in a standardized way across different faculties and/or curricula. In addition, MCQs provide high reliability and objectiveness. One testing format that utilizes the virtues of multiple choice items is progress testing. The main aim of this procedure is to follow students learning trajectories over the course of their studies. Progress tests depict the development – hence '*progress'* – of knowledge over the course of (medical) education. To this aim, students are repeatedly tested with different items while the overall distribution of content (topics/subjects) remains constant. The items are intended to assess knowledge relevant for a doctor's "first-day-in-practice" and thus are comparable to the demands of a graduation exam. In case of the Progress Test Medicine (PTM) of the *Charité Universitätsmedizin Berlin* each question is reviewed twice by a multidisciplinary review team with regard to this aspect. Further details on progress test can be found in Wrigley et al. (Wrigley et al., 2012).

Typically, progress tests are organized as inter-institutional or even international collaborations (Freeman et al., 2010; Tio et al., 2016). Such cooperative efforts provide an optimal basis for comparisons of large cohorts of students' levels of medical knowledge across institutions. In Germany, the PTM is set up as an inter-institutional cooperation. The test is conducted biannually (once every semester) and consists of 200 multiple choice questions which are constructed as single best answer items (Nouns and Georg, 2010). Students from their 1st to the 6th academic years take part in the test. A total number of about

180.000 participants from 17 medical schools in Germany and Austria have taken the test since its introduction in 1999. As the PTM is a formative assessment tool, students do not prepare extensively. Consequently, results can be assumed to be unbiased by test-preparation efforts and thus reflect students' readily retrievable knowledge. Previously, results from the PTM have been used for curricular comparisons and benchmarking the performance of students (in different subjects) from different backgrounds (home medical school, types of curricula) (Nouns et al., 2012; Schauber et al., 2015; Tio et al., 2016; Verhoeven et al., 1998). Three specific features of the PTM make this an attractive assessment tool for our question delineated above. First, the PTM is synchronously administered across many medical schools in Germany. Second, each PTM contains between 15 and 20 questions testing anatomy knowledge. Third, students participate at all stages of their studies. Hence, information about medical students' performance in their 5th and 6th academic years can be extracted from the PTM and used for assessing their knowledge in human anatomy across several medical faculties in Germany.

In order to approach the question whether or not students' performances are satisfactory, an objective, reliable and valid standard is crucial. While an assessment tool such as the PTM provides data on *actual* performance levels, these empirical results need to be compared to what teachers, lecturers, or experts (i.e., content matter experts) *expect* from graduates. If such expectancies are obtained in a systematic and objective manner, they can serve as a reference– a *standard*– to which students' actual performances can be compared to. Put briefly, such standards set by content matter experts define how many students *should* correctly respond to a specific question. The procedures for obtaining such a standard for a whole set of questions (i.e., a test) are referred to as "standard setting procedures" in the educational assessment literature. When making claims about performance or ability deficits, standards are crucial as they link the tested content to the expected competence levels (Bandaranayake, 2008; Ben-David, 2000).

Standard-setting procedures have already been used for establishing references for progress test results (Verhoeven et al., 2002) as well as for the evaluation of anatomy knowledge (Prince et al., 2005). One of the best-known method, the Angoff method, is suitable for setting criterion referenced standards for multiple choice examinations (Angoff, 1971; Bandaranayake, 2008). According to this method a number of judges have to estimate for each question the percentage of a group of minimally-competent candidates who are at the borderline of pass and fail (i.e., a 'borderline'-examinee or minimally-competent student) that would respond correctly. For instance, the judges have to answer a question such as: "How often would you expect a group of students on the verge of passing/failing this exam to

answer this specific item correct?" However, research on this procedure has found that estimating the responses of a group of ‘‘borderline'-examinees is difficult for unexperienced judges (Norcini, 1994). Additionally, the estimation of the percentage of correct responses of a group is a problem, as even experienced judges tend to choose values between 40% and 60%. Thus, Impara and Plake (1997) proposed a modified Angoff-procedure, also referred to as the he 'Yes/No-Method'. In this variant, judges have to imagine *one* 'borderline'-candidate and to decide for each question, if he/she would give a correct answer or not (Chinn and Hertz, 2002). The reference standard is then calculated as average number of questions a 'borderline'-examinee is expected to answer correctly by the judges.

In summary, to investigate the question whether a deficit in medical students' anatomical and topographical knowledge can be found empirically, two sources of information are needed: First, medical students' retrieval of anatomical content has to be assessed thoroughly. Second, any objective standard for comparison needs to link the actually tested content to an expected level of competence (Bandaranayake, 2008; Ben-David, 2000). Importantly, this question should be addressed not just locally but rather in a multi-centre study. Both students' performances and content matter experts ratings need to be obtained from different medical schools in order to reduce the influence of locally varying standards, and/or curricular-specific effects.

In the present study we aimed at answering the question whether German medical students' knowledge of human anatomy is sufficient in order to enable curriculum planners to account for potential deficits. Therefore we compared the performance of upper year students on anatomical questions of the Berlin PTM with a standard established by a modified Angoff procedure.

Material and Methods

Progress test and participants

The Berlin Progress Test Medicine (PTM) was used to assess students' retrieval of anatomy content at the end of their studies. Therefore, results of medical students in their academic years five and six from three consecutive tests were included in this study. Data was obtained from ten medical schools which conducted the PTM in Germany in years 2008 and 2009. The numbers of participating students were N=1470, N=1951, N=1962 for the three tests taking place in October 2008, April 2009, and October 2009, respectively. While the majority of the students participating (3556, 66.1%) were from regular curricula, 1827

(33.9%) of the students were from reformed curricula. Table 1 lists the number of students from each university and curriculum type for all three tests.

In general, 200 test items in one progress test are randomly sampled from an item pool containing approximately 5000 multiple choice items. All questions which are used in a particular PTM are excluded from the sampling procedure for four consecutive tests (i.e. for two years) in order to reduce recognition effects. Questions in the PTM are classified into subjects and organ systems. Subjects include basic sciences (anatomy - including gross anatomy, histology, and embryology -, biochemistry and chemistry, physiology and physics) as well as clinical disciplines (internal medicine, surgery, paediatrics, gynaecology, psychiatry, ophthalmology, dermatology, emergency medicine, orthopaedics, family medicine, urology, and neurology). Organ systems include skin, immune system, endocrine system, musculoskeletal system, respiratory system, digestive system, urinary system, reproductive system, nervous system and sensory organs. All items are administered in single-best answer multiple-choice format and typically make use of clinical vignettes (i.e., case or patient descriptions). Before test administration, independent expert committees have judged each single item with regard to its content and formal appropriateness. In a post-test quality assurance procedure, questions which either show non-optimal psychometric properties (e.g., extreme difficulties or negative correlation with the total score) or flagged up by students as potentially erroneous are again submitted to expert panels for final review and approval (or disapproval).

As common in progress tests, students are able to choose a “don’t know” option to “dismiss” content they cannot readily answer. This is necessary because students in their 1st academic years cannot answer a substantial part of the questions. As the test score is obtained by negative marking of incorrect answers, that is, the number of wrong answers is subtracted from the number of correct answers, without this option lower year students would have to guess answers to many items, thereby introducing a strong random component and a likely negative bias into their final score. The “don’t know” option thus removes the pressure from all students, including those in their upper years, to make any unnecessary random choices, leading to a more reliable assessment of their actual knowledge. Although no pass or fail decisions are made on the basis of the test results, it is compulsory for students to take the test twice a year.

Anatomy questions used for standard setting

In the PTM item pool, 155 questions are assigned directly to the subject field of “anatomy/biology”. However, there are about 40 additional questions which focus on

anatomical knowledge, typically in a clinical context such as surgery. In order to collect sufficiently large set of questions for a standard setting, anatomy-relevant MCQs from three consecutively conducted PTMs were pooled. The pool initially contained 49 questions. From this pool, we selected a final set of 33 questions. The selection was made by two experienced anatomists. Questions were excluded on the basis of (1) similarity to other questions and (2) over-representation of a certain anatomical theme, such as organ systems. In the final set all organ systems were covered as evenly as possible (Suppl. Table1).

The 33 questions were subsequently classified into 3 different categories for more fine-grained analysis by three experienced anatomists (all from the Charité, Berlin) independently. The categories were as follows: (1) questions testing only factual anatomical knowledge (e.g. identifying branches of a given nerve), (2) questions testing the ability to reproduce anatomical structure-function relationships (e.g. identifying movements served by a given muscle or muscle group, "simple application"*)*, and (3) questions where anatomical knowledge has to be applied in a clinical context (for example questions referring to a case history, "clinical application").

Judges for standard setting panels

Judges were recruited from anatomical institutes of five German university medical schools. All of these medical schools conduct the PTM and results of their students were included. The five panels together comprised 31 anatomists of whom 15 members graduated in human medicine, 11 in biology and 5 in other subjects. The teaching experience of the judges varied from below five years to more than 30 years. Published criteria for the selection require that a judge had been teaching anatomy to medical students for at least two years in order to become a member of a panel. In fact, anatomists with sufficiently long teaching experience at least meet three recommended criteria for standard setting panellists (Ben-David, 2000): (1) subject-specific expertise, (1) interest in medical education and (3) experience in examination methods. Familiarity with the level of the candidates was assured by confronting judges with actual test results during the standard setting procedure (see also Standard setting method below). The recommended criterion of good problem solving skills is difficult to monitor in the process of selecting panellists.

Standard setting method

The standard setting procedure was conducted independently across the five panels at the participating institutions. Each panel met for one session (approximately three hour long). One week before the meeting, panel members were informed about the project, the specific schedule of the standard setting and received literature via email. Each meeting was

moderated by the same facilitator and followed a standardized schedule. The '*Yes/No-Method*' (Chinn and Hertz, 2002; Impara and Plake, 1997), a modified Angoff procedure, was chosen to establish a criterion referenced standard for the assessment of the students' performance on anatomy questions in the PTM. Panel members were asked to imagine a 'borderline'-examinee after completing his/her 4th year of studies and to judge whether the examinee would answer the questions correctly.

Confronting judges of standard setting panels with actual test results has been shown to improve the correlation of the judges' estimations with the actual item difficulties (Norcini et al., 1988). For this reason and because estimating a borderline-examinee's responses is difficult (Norcini, 1994), in particular for panellists who are not experienced in standard setting procedures, the judges were subsequently confronted with actual test results: After a first round of evaluation of the question set, panel members were asked to focus on three particular items from the set which had well-defined difficulty levels, designated as easy, intermediate and difficult by three independent experts from the Charité, anatomists with a long teaching experience. The difficulty levels of these particular items were compared to the performance of the students on these three questions. Afterwards, the judges had the opportunity to revise their evaluation of the question set. Panel meetings were closed with a final evaluation. The final score was then calculated as the average number of questions a theoretical 'borderline'-examinee would be expected to correctly answer as estimated by the judges. In this study a common standard covering all questions used and specific standards for each question category were calculated. The scores were then used to decide if medical students' performance was sufficiently good or not.

Statistical evaluation

Data are presented as means $\pm$ SD and a significance level of $p \leq 0.05$ was applied throughout this study. Statistical significances were calculated using the Kruskal-Wallis H test.

Ethical approval

The study has been approved by a research ethics committee of the *Charité – Universitätsmedizin Berlin*.

Results

Standard setting score

Five independent panels of a total of 31 anatomists from five German medical universities established a standard (cutting score) for 33 anatomy questions selected from three consecutive rounds of the PTM (winter semester 2008/2009 – winter semester 2009/2010). The selection of the items and the standard setting procedures were conducted in spring 2010, after the three progress test had taken place. The cutting score corresponded to the estimated percentage of correctly answered questions by a student with minimal acceptable performance ("borderline").

The mean common standard setting score was 60.4%, corresponding to an average of 19.9 correctly answered questions from the set used in the study (Tab. 2). As there were no significant differences (Kruskall-Wallis H-test p=0.731) in the standards established by the five independent panels (Tab.2), a common standard was calculated and used in the assessment of the students' performance. To check whether the standard setting outcome varied as a function of expertise, mean estimated scores were calculated after separating the judges into groups according to their degree subject or teaching experience, however these scores did not show any significant difference either (Suppl. Tab. 2 and 3) confirming that the common standard was valid and also robust.

During the standard setting procedure judges were confronted with actual test results of students in three items with well-defined difficulty levels and were then given the opportunity to revise their decision for each of the 33 items used in the standard setting. During this procedure, the judgement for an average of 3.3 items was changed, ranging from 0 (6 panellists) to a maximum of 7 (2 panellists) (Suppl. Tab. 4). In 60 cases an item-decision was changed from "no" (meaning a borderline examinee would not be expected to answer correctly) to "yes" (meaning a borderline examinee would be expected to answer correctly), in 42 cases from "yes" to "no". After these corrections the standard setting score was increased minimally from an initial value of 58.6% to the final standard of 60.4%.

Comparison of students' progress test results with standard setting scores

Students with solid theoretical knowledge and good skills in applying this knowledge are expected to perform well above the cutting score. Results of medical students in their fifth and sixth year, corresponding to the last year of their clinical science studies and their clinical placement, from 10 German medical universities were collected for evaluation. However, the average score of the students on the anatomical questions included in this study was substantially lower at 29.9%, corresponding to only 10 correct answers out of the 33 on

average (Tab. 3). Thus, the students scored significantly (Kruskall-Wallis H-test p<<0.001) below the standard, indicating that their knowledge of human anatomy is underneath the expected minimally sufficient performance. The same results were obtained when a standard calculated after the first round of estimations was used for comparison (Kruskall-Wallis H-test p<<0.001). Results from a generalizability study suggest that the findings are stable across the involved universities and the different PTMs (Tab. 4). Indeed, the two estimated variance components accounted for only about 0.5% of the total variance, indicating an overall reliability $G_{(relative)}$=0.89. Performance of the students on the anatomical MCQs correlated well to their overall performance in the respective PTM (Suppl. Table 5). The actual item difficulties of each individual MCQ are provided in Suppl. Tab. 6.

Standards and students' results in different categories of questions

Our results indicate that the students' overall performance was significantly below the standard. However, it remained unclear whether this was equally true for all types of multiple choice items, namely those testing (1) factual knowledge, (2) simple application or (3) clinical application of anatomical knowledge (see Methods). Therefore, we compared the students' results in these three categories to the estimated cutting scores and found that for all categories the performance was lower than expected. However, the deficits in the performance were not uniformly distributed across these categories (Tab. 5, Kruskall-Wallis H-test p=0.004): while for factual knowledge the students' score was 39.7% below the expected score (40.8% vs. expected 67.7%; Tab. 5, Kruskall-Wallis H-test p=0.015), for clinical application their performance was 58% lower (22.5% vs. expected 53.5%, Tab. 5, Kruskall-Wallis H-test p<0.001). Taken together, the data indicate that German medical students' knowledge of human anatomy is lower than a standard set by independent committees of content matter experts. Furthermore, their deficits are the strongest for application of anatomical knowledge in a clinical context, for the category which has the highest relevance for their future work as medical practitioners.

Discussion

This study was performed with the aim to assess presumed deficits in the anatomy knowledge of upper-year medical students at German university medical schools. Our results revealed that students in the 5th and 6th year of their studies performed considerably below the established standard in a progress test, confirming that there are deficits in their factual anatomical and topographical knowledge, as often claimed by members of surgical

disciplines. Furthermore, our results demonstrated that the students had particular difficulty in applying their knowledge to problems in the context of a clinical setting.

Our result that upper-year students performed considerably below the established standard in the PTM depends critically on the choice of the assessment tool. Therefore we need to analyze whether the MCQs borrowed from the biannual PTM are adequate for our evaluation purpose. MCQs can be used to test knowledge ("knows") and competence ("knows how") described in Miller's pyramid of competence (Miller, 1990). Complex spatial orientation, which is in particular an important aspect of anatomy and its application during clinical practice like surgery, contributes to different levels of the pyramid such as competence, performance ("shows how") and action in clinical practice ("does") and cannot be properly assessed by MCQs. Commonly assessment tools such as objective structured practical examination or "*anatomy spot test*" on prosected cadavers are better suited for testing three-dimensional, topographical orientation (Rowland et al., 2011; Schoeman and Chandratilake, 2012). Within these limitations, MCQs are appropriate for testing anatomy knowledge, which has mainly either factual or conceptual quality. In fact, previous studies demonstrated that students' performance was better when their anatomical knowledge was tested by MCQs in comparison to that when other assessment tools were used (Bergman et al., 2011; Hobsley, 1976; Moqattash et al., 1995) suggesting that MCQs offer a robust but benign form of testing of anatomical knowledge. Therefore, the students' poor performance on MCQs from the PTM is unlikely to be attributable to the format of the test chosen for evaluation.

The validity of the standard strongly depends on the themes covered by the MCQs used. A limitation of our study may, thus, lie in an uneven representation of the diverse topics of anatomy, such as the different organ systems in the question set. In selecting the MCQs from the original pool, we aimed at eliminating any such bias. Although, the limited number of questions did not permit to cover all fields of anatomy in a fully even manner, the size of the question set used in our study (a total of 33 MCQs) is large enough to achieve an unbiased assessment and provides representative test results for the fields covered as shown by the generalizability study we performed (see Table 4). Accordingly, a small deviation in the performance of the students from the standard might be explicable by such inhomogeneity in the question set, however the performance of the students observed in our study is markedly lower than the expected cutting score (29.9% vs. 60.4%) indicating a substantial deficit in their knowledge and supporting our main conclusion.

A further aspect we need to consider is whether the standard established in this study is appropriate for our evaluation purpose. One could argue that anatomists as experts in their

field set standards which are higher than appropriate. However, being an expert in the field of the exam is one of five major criteria recommended for the selection of the judges in standard settings (Ben-David, 2000). Indeed, there is a broad agreement on the need of subject specific expertise as distinguishing feature of judges invited to establish the standard setting (Cusimano, 1996; Jaeger, 1991). Anatomists with teaching experience meet two more of the recommended criteria: (1) they are intimately interested in medical education and (2) familiar with examination methods. In fact, a study comparing standards on anatomical knowledge established by panels of different subgroups of panellists (anatomists, clinicians, graduates, and 4-year medical students) revealed that the standard established by the anatomist was nearest to the mean of all panellists estimates in one study (Prince et al., 2005). However, an aspect which may constitute a problem is the judges' understanding of the students` level of knowledge at the end of their studies, as most of the anatomy teaching takes place in the early years. On the other hand even if anatomy is taught in the first years, teachers' expectations are defined by a concept of what a (young) doctor needs as background anatomical knowledge in everyday clinical practice. In this respect, the upper-year student cohorts assessed in our test dramatically fail these expectations. On the other hand it would be interesting to investigate in the future if a standard established by panels of surgeons, who are experts in anatomy from a clinical point of view, differs from the standard established by anatomists. Furthermore, a combination of anatomists, experienced clinicians and graduates in a standard setting panel could give a better reference that needs to be tested in the future.

For our standard setting procedure we decided to confront the panellists with actual test results from students before a second round of evaluation. This kind of procedure has been shown to improve the correlation of the judges' estimations with the actual item difficulties (Norcini et al., 1988). In a strict sense this approach makes the standard not completely criterion referenced, as it implicates a feedback element from the performance of the test group. However, imagining a borderline-examinee is difficult in particular for persons who are not experienced in standard setting procedures (Norcini, 1994). Furthermore, the estimation of a borderline performance includes both aspects, the knowledge content and the average student performance. Therefore, feeding back the actual test results to the standard seems to be a reasonable procedure in order to avoid estimates far from reality. In our study judges were only confronted with actual results of three items, thus deviating only slightly from the pure criterion-referenced approach. In fact, in our study the standard calculated from estimations before the confrontation with actual results was very similar to the standard

calculated after the confrontation and the performance of the students was considerably below both of these standards.

Another aspect to be discussed regarding the standard setting procedure is that panel members were asked to imagine a 'borderline'-examinee after completing his/her 4th year of studies, although the MCQ of the PTM are designed to assess knowledge needed by a doctor "first day in practice". The test itself can be conducted beginning from the first semester to gain information on the increase of knowledge throughout the complete studies. Our study aimed at revealing/excluding potential deficits in the anatomical knowledge of medical students in the final years of their studies. Therefore the judges were requested to estimate the level of a student after his/her 4th year.

The high relevance of anatomy for clinical practice (Arraez-Aybar et al., 2010; Hofer, 2006; Older, 2004) ensures the subject a central position in medical curricula. In the last decades, undergraduate medical education underwent a significant change world-wide (Drake, 2014) with new methods for teaching and learning introduced (Gunderman and Wilson, 2005; Reidenberg and Laitman, 2002). Integration of disciplines and introduction of problem based learning were two central elements of these curricular innovations. However, these developments were accompanied by a marked reduction in the amount of anatomy classes in the curricula (Drake et al., 2009). Although no clear relationship between the didactic approach of a medical school and the level of anatomical knowledge of students has been observed in previous comparisons (Prince et al., 2005), the total time spent on teaching of anatomy, anatomical facts in clinical context and the recurrence of anatomical themes in upper years of medical curriculum have been reported to be important factors (Bergman et al., 2008). In particular, teaching time for dissection of cadavers has been declining over the past years (Gartner, 2003). Some reformed curricula (McLachlan, 2004; McLachlan et al., 2004) including the Reformed Medical Curriculum of the Charité had completely removed dissection from the program. Indeed, students from that reformed curriculum self-estimated their anatomy knowledge as worse than students from the traditional curriculum, when asked if their knowledge is sufficient for clinical practice. Moreover, their performance in anatomical PTM questions was significantly worse compared to the performance of students from the traditional curriculum (Brunk et al., 2013). In good agreement, our results now clearly indicate an insufficient anatomical knowledge of recent upper-year medical students at German universities. However, more in depth analysis of the PTM data would be required to evaluate whether a tighter relationship between total teaching time at the different medical schools and the knowledge of the students exist. Finally, one additional relevant aspect is the

representation of anatomical content throughout different curricula. If differences between the medical schools correlated with different results in PTM questions exist, this information would represent an important argument in favour or against the decision of revising curricula in direction of teaching anatomy longitudinally.

Recent concerns of members of surgical disciplines regarding the anatomy knowledge of upper-year medical students' and young doctors could either be consequences of deficits in factual knowledge or problems in application of this knowledge in the daily clinical practice. Our results reveal that unfortunately both mechanisms contribute unfavourably to this problem: scoring 40% below the expected minimally sufficient standard, our test indicate a substantial deficit in the factual knowledge of the students. Moreover, an even worse deficit of 58% for questions assessing the clinical application of the knowledge reflects a grave problem beyond the mere reproduction of facts. Thus our results support the opinion, that teaching anatomical facts in clinical context is crucial (Bergman et al., 2011). This should include the vertical integration of the subject throughout medical curricula with the aim of enhancing retention by recurring to anatomical content in the teaching of clinical disciplines. Given the fact that biomedical knowledge has been shown to be important for making diagnoses (Woods, 2007; Woods et al., 2007a, b; Woods et al., 2006) the methods, the curricular time-points and the frequency of (anatomy) teaching have to be chosen with regard to their effectiveness and sustainability in promoting clinical application.

Conclusions

There is an ongoing dispute about the impact of anatomy knowledge of medical students and young doctors on patient safety (Collins, 2009; Yammine, 2014). In connection with the concerns raised by members of surgical disciplines and the results obtained in our study it seems to be essential to review the time and the way anatomy is taught to medical students. In particular, our finding that the clinical application is a major problem has to be urgently addressed. Strategies for facilitating transfer of basic science contents to clinical problems suggested previously include problem solving from the beginning of learning and the use and comparison of different examples for identifying underlying similar or dissimilar concepts (Norman, 2009).

In this context a highly relevant, but seemingly paradoxical finding is that a negative correlation appears to exist between biomedical knowledge and the acquisition of clinical knowledge (Schauber et al., 2013) despite the broad agreement of experts regarding the high relevance of doctors' biomedical background for clinical practice (Boshuizen and

Schmidt, 1992; Rikers et al., 2005b; Woods et al., 2006). As this study is mainly based on data from students from traditional curricula (66.1%), which offer largely separated basic science and clinical education, our results indicate that within traditional curricula the clinical application of anatomy knowledge is not facilitated sufficiently. Possibly a model of integrated teaching (horizontal integration with teaching anatomy in a clinical context and vertical integration by repetition of anatomy contents at different time points of a curriculum) could be a beneficial approach. Taken together, we propose that the integrated approach to anatomy teaching in early years of the curriculum and the systematic recapitulation of anatomical facts in clinical context in later years should be promoted and the trend of continuous reduction in total teaching time should be critically reviewed when revising and developing new medical curricula.

Aknowledgements

The authors thank Imre Vida for valuable discussions and comments on the manuscript.

Tables and figures

Table 1

University	Semester					Curriculum		Total
	9	10	11	12	13	traditional	reformed	
1	820	835	135	7	23	1522	298	1820
2	260	195	200	0	0	1	654	655
3	372	181	374	0	0	734	193	927
4	365	261	15	0	0	0	641	641
5	329	2	100	0	0	431	0	431
6	202	186	3	0	0	390	1	391
7	149	54	0	0	0	203	0	203
8	30	9	2	0	0	1	40	41
9	0	73	156	0	0	229	0	229
10	24	7	14	0	0	45	0	45
Total	2551	1803	999	7	23	3556	1827	5383

Table 1: Number of participating students per semester, university and type of curriculum.

Table 2

	Score (%)	SD (%)	Statistical significance
University A	62.3	10.6	0.731
University B	59.1	9.2	
University C	61.9	9.7	
University D	56.1	13.2	
University E	62.6	2.9	
Common score	60.4	10.5#	

Table 2: Common and university-specific standard-setting scores. The scores were calculated as the mean values of the judges individual estimates of the percentage of correctly answered questions by a borderline student with minimally sufficient knowledge in anatomy for the independent panels (Universities A - E) and subsequently for all judges (Common score). To assess the statistical significance of differences among the expert panels Kruskal-Wallis H test was used; the high P value indicates no statistical difference justifying the pooling of the estimates and the calculation of a common score

Table 3

	Score (%)	SD (%)	Statistical significance
Cutting score according to standard setting	60.4	23.6#	<<0.001*
Students' results	29.9	14.6	

Table 3: Comparison of the cutting score established in the standard setting and the students' performance on the anatomical multiple choice items from the PTM. Kruskal-Wallis H test was applied to statistically evaluate the difference in the scores.

Note that the SD value displayed in this table differs from that of the common score in Table 1, because here the arbitrary item difficulty derived from the standard setting was used for calculation of the mean and the SD, whereas in Table 1 the variance between the individual scores estimated by the different judges was compared. While the means converge, SD differ in both value and the conceptual relevance for these two comparisons.

Table 4

Facet	Variance	N	
Student	0.04	10981	
Items	0.02	11	
Occasion	0.00	3	
Residual	0.15		
Reliability			0.89 (relative)

Table 4: The results from the generalizability study show an overall reliability on 0.89

Note: The between-university variance was estimated as 0.00 and therefore excluded from the further analysis.

Table 5

Question category	Number of questions	Cutting scores		Students' results		Statistical significance
		Score (%)	SD (%)	Score (%)	SD (%)	
Factual knowledge	9	67.7	20.9	40.8	14.9	0.015*
Simple application	5	73.5	19.9	38.3	7.2	0.028*
Clinical application	19	53.5	23.2	22.5	11.1	<<0.001*
Statistical significance		0.119		0.004*		

Table 5: Comparison of the cutting scores and the students' performance in the PTM for the 3 different question categories. Note that the students scored significantly below the standard for all 3 category but performed particularly bad on questions requiring clinical application of their knowledge.

Kruskal-Wallis H test was used to calculate the statistical significance of observed differences between and within the item categories.

Supplementary data

Supplementary Table 1

Theme	Number / % of questions in PTM 19-21	Number / % of questions for standard setting
General anatomy	-	-
Cell biology / histology	3 / 6.1 %	3 / 9.1%
Embryology	4 / 8.2 %	3 / 9.1%
Physical examination	3 / 6.1 %	3 / 9.1%
Skin	-	-
Immune system	-	-
Endocrine system	1 / 2.0 %	1 / 3.0%
Head	2 / 4.1 %	2 / 6.1%
Upper limb	5 / 10.2%	3 / 9.1%
Lower limb	9 / 18.4%	5 / 15.2%
Inguinal canal	3 / 6.1 %	1 / 3.0%
Heart	4 / 8.2 %	2 / 6.1%
Airways and lung	-	-
Digestive system	2 / 4.1 %	2 / 6.1%
Urinary system	-	-
Reproductive system	4 / 8.2 %	3 / 9.1%
Peripheral nervous system	6 / 12.2%	3 / 9.1%
Central nervous system	4 / 8.2 %	2 / 6.1%
Sensory organs	-	-
Total	49 / 100%	33 / 100%

Supplementary Table 1: Organ system specific distribution of multiple choice items testing anatomy knowledge in the PTM in October 2008, April 2009, and October 2009 and in the questions set selected for the standard setting. Several questions from the PTM were excluded in order to avoid overlap of the questions and to achieve an even representation across the organ systems.

Supplementary Table 2

graduation	Number of judges	Score (%)	SD (%)	Statistical significance
Human medicine	15	58.0	9.7	0.170
Biology	11	61.4	12.1	
Others	5	64.2	7.3	

Supplementary Table 2: Mean standard-setting scores calculated based on the degree subjects of judges. The score corresponds to the estimated percentage of correctly answered questions by a borderline student with minimally-sufficient knowledge in anatomy. Statistical significance between the groups was assessed using the Kruskal-Wallis H test but it revealed no significant difference.

Supplementary Table 3

Teaching experience (years)	Number of judges	Score (%)	SD (%)	Statistical significance
< 5	4	49.2	3.3	0.280
5 - < 10	7	64.1	12.2	
10 - < 20	8	59.8	9.8	
20 - < 30	6	59.1	8.2	
> 30	6	64.6	9.0	

Supplementary Table 3: Mean standard-setting scores calculated according to the teaching experience of judges (in years). The score corresponds to the estimated percentage of correctly answered questions by a borderline student with minimally-sufficient knowledge in anatomy. Statistical significance testing using the Kruskal-Wallis H test revealed no significant difference.

Supplementary Table 4

Number of revised items	0	1	2	3	4	5	6	7
Number of participants	6	1	3	8	2	5	4	2

Supplementary Table 4: Frequency of revised decisions in the second stage of the standard setting procedure. The number of revision varied from 0 to 7 with a median of 3.

Supplementary Table 5

Semester	Students with > 10 correct answers in the PTM	Students with > 50 correct answers in the PTM	Students with > 100 correct answers in the PTM
8	11,78	14,9	18,83
9	10,95	14,27	18,19
10	11,77	15,35	19,7

Supplementary Table 5: Number of correctly answered anatomy question in different PTM performance groups. Note that the number of correct answers to anatomy questions increases with the performance for all 3 semesters, but shows no correlation with the semesters.

Supplementary Table 6

Factual knowledge	0.376
	0.534
	0.473
	0.586
	0.489
	0.089
	0.227
	0.424
	0.478
Simple application	0.475
	0.347
	0.297
	0.331
	0.462
Clinical application	0.046
	0.291
	0.117
	0.125
	0.187
	0.131
	0.108
	0.395
	0.274
	0.300
	0.236
	0.267
	0.312
	0.405
	0.166
	0.159
	0.051
	0.385
	0.327

Supplementary Table 6: Actual item difficulties of anatomy-relevant questions used in the standard setting. Values were calculated as the proportion of students giving correct answers to the individual questions.

References

Angoff, W., 1971. Scales, norms and equivalent scores. American Council on Education, 508-600.

Arraez-Aybar, L.A., Sanchez-Montesinos, I., Mirapeix, R.M., Mompeo-Corredera, B., Sanudo-Tejero, J.R., 2010. Relevance of human anatomy in daily clinical practice. Annals of anatomy = Anatomischer Anzeiger : official organ of the Anatomische Gesellschaft 192, 341-348.

Bandaranayake, R.C., 2008. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. Med Teach 30, 836-845.

Ben-David, M.F., 2000. AMEE Guide No. 18: Standard setting in student assessment. Med Teach 22, 120-130.

Bergman, E.M., Prince, K.J.A.H., Drukker, J., van der Vleuten, C.P.M., Scherpbier, A.J.J.A., 2008. How Much Anatomy Is Enough? Anat Sci Educ 1, 184-188.

Bergman, E.M., van der Vleuten, C.P., Scherpbier, A.J., 2011. Why don't they know enough about anatomy? A narrative review. Med Teach 33, 403-409.

Bergman, E.M., Verheijen, I.W., Scherpbier, A.J., Van der Vleuten, C.P., De Bruin, A.B., 2014. Influences on anatomical knowledge: The complete arguments. Clin Anat 27, 296-303.

Biasutto, S.N., Caussa, L.I., Criado del Rio, L.E., 2006. Teaching anatomy: cadavers vs. computers? Annals of anatomy = Anatomischer Anzeiger : official organ of the Anatomische Gesellschaft 188, 187-190.

Boshuizen, H.P.A., Schmidt, H.G., 1992. On the Role of Biomedical Knowledge in Clinical Reasoning by Experts, Intermediates and Novices. Cognitive Sci 16, 153-184.

Brunk, I., Georg, W., Schauber, S., 2013. The state of students' knowledge about human anatomy within two different medical curricula. Poster at AMEE conference

Cahill, D.R., Leonard, R.J., Marks, S.C., Jr., 2000. A comment on recent teaching of human anatomy in the United States. Surgical and radiologic anatomy : SRA 22, 69-71.

Chinn, R.N., Hertz, N.R., 2002. Alternative approaches to standard setting for licensing and certification examinations. Appl Meas Educ 15, 1-14.

Chirurgie, D.G.f., 2009. Defizite bei Anatomie-Kenntnissen gefährden Patientenversorgung. Pressemitteilung Deutsche Gesellschaft für Chirurgie http://presseservice.pressrelations.de/standard/result_main.cfm?aktion=jour_pm&r=364400&quelle=0&pfach=1&n_firmanr_=121968&sektor=pm&detail=1#.

Collins, J.P., 2009. Are the changes in anatomy teaching compromising patient care? The Clinical Teacher 6, 18-21.

Cusimano, M.D., 1996. Standard setting in medical education. Academic medicine : journal of the Association of American Medical Colleges 71, S112-120.

Drake, R.L., 2014. A retrospective and prospective look at medical education in the United States: trends shaping anatomical sciences education. Journal of anatomy 224, 256-260.

Drake, R.L., McBride, J.M., Lachman, N., Pawlina, W., 2009. Medical education in the anatomical sciences: the winds of change continue to blow. Anat Sci Educ 2, 253-259.

Freeman, A., Van der Vleuten, C., Nouns, Z., Ricketts, C., 2010. Progress testing internationally. Med Teach 32, 451-455.
Gartner, L.P., 2003. Anatomical sciences in the allopathic medical school curriculum in the United States between 1967-2001. Clin Anat 16, 434-439.
Gunderman, R.B., Wilson, P.K., 2005. Exploring the Human Interior: The Roles of Cadaver Dissection and Radiologic Imaging in Teaching Anatomy. Acad Med 80, 745-749.
Hobsley, M., 1976. Assessment of anatomy in the Primary FRCS. Annals of the Royal College of Surgeons of England 58, 382-384.
Hofer, M., 2006. Potential improvements in medical education as retrospectively evaluated by candidates for specialist examinations. Dtsch Med Wochenschr 131, 373-378.
Impara, J.C., Plake, B.S., 1997. Standard setting: An alternative approach. J Educ Meas 34, 353-366.
Jaeger, R., 1991. Selection of judges for standard setting. Education measurement Issues and Practice 10, 3-6.
McLachlan, J.C., 2004. New path for teaching anatomy: living anatomy and medical imaging vs. dissection. Anatomical record. Part B, New anatomist 281, 4-5.
McLachlan, J.C., Bligh, J., Bradley, P., Searle, J., 2004. Teaching anatomy without cadavers. Med Educ 38, 418-424.
Miller, G.E., 1990. The assessment of clinical skills/competence/performance. Acad Med 65, S63-67.
Moqattash, S., Harris, P.F., Gumaa, K.A., Abu-Hijleh, M.F., 1995. Assessment of basic medical sciences in an integrated systems-based curriculum. Clin Anat 8, 139-147.
Mylopoulos, M., Woods, N., 2014. Preparing medical students for future learning using basic science instruction. Med Educ 48, 667-673.
Norcini, J.J., 1994. Research on Standards for Professional Licensure and Certification Examinations. Eval Health Prof 17, 160-177.
Norcini, J.J., Shea, J.A., Kanya, D.T., 1988. The Effect of Various Factors on Standard Setting. J Educ Meas 25, 57-65.
Norman, G., 2009. Teaching basic science to optimize transfer. Med Teach 31, 807-811.
Nouns, Z., Schauber, S., Witt, C., Kingreen, H., Schuttpelz-Brauns, K., 2012. Development of knowledge in basic sciences: a comparison of two medical curricula. Med Educ 46, 1206-1214.
Nouns, Z.M., Georg, W., 2010. Progress testing in German speaking countries. Med Teach 32, 467-470.
Older, J., 2004. Anatomy: A must for teaching the next generation. Surg-J R Coll Surg E 2, 79-90.
Prince, K.J., Scherpbier, A.J., van Mameren, H., Drukker, J., van der Vleuten, C.P., 2005. Do students have sufficient knowledge of clinical anatomy? Med Educ 39, 326-332.
Prince, K.J., van Mameren, H., Hylkema, N., Drukker, J., Scherpbier, A.J., van der Vleuten, C.P., 2003. Does problem-based learning lead to deficiencies in basic science knowledge? An empirical case on anatomy. Med Educ 37, 15-21.
Pryde, F.R., Black, S.M., 2005. Anatomy in Scotland: 20 years of change. Scottish medical journal 50, 96-98.

Reidenberg, J.S., Laitman, J.T., 2002. The new face of gross anatomy. The Anatomical record 269, 81-88.
Rikers, R.M., Loyens, S., te Winkel, W., Schmidt, H.G., Sins, P.H., 2005a. The role of biomedical knowledge in clinical reasoning: a lexical decision study. Acad Med 80, 945-949.
Rikers, R.M.J.P., Schmidt, H.G., Moulaert, V., 2005b. Biomedical knowledge: Encapsulated or two worlds apart? Appl Cognitive Psych 19, 223-231.
Rowland, S., Ahmed, K., Davies, D.C., Ashrafian, H., Patel, V., Darzi, A., Paraskeva, P.A., Athanasiou, T., 2011. Assessment of anatomical knowledge for clinical practice: perceptions of clinicians and students. Surgical and radiologic anatomy : SRA 33, 263-269.
Saltarelli, A.J., Roseth, C.J., Saltarelli, W.A., 2014. Human cadavers Vs. multimedia simulation: A study of student learning in anatomy. Anat Sci Educ 7, 331-339.
Schauber, S.K., Hecht, M., Nouns, Z.M., Dettmer, S., 2013. On the role of biomedical knowledge in the acquisition of clinical knowledge. Med Educ 47, 1223-1235.
Schauber, S.K., Hecht, M., Nouns, Z.M., Kuhlmey, A., Dettmer, S., 2015. The role of environmental and individual characteristics in the development of student achievement: a comparison between a traditional and a problem-based-learning curriculum. Adv Health Sci Educ 20, 1033-1052.
Schoeman, S., Chandratilake, M., 2012. The anatomy competence score: a new marker for anatomical ability. Anat Sci Educ 5, 33-40.
Sinclair, D., 1975. The two anatomies. Lancet 1, 875-878.
Tio, R.A., Schutte, B., Meiboom, A.A., Greidanus, J., Dubois, E.A., Bremers, A.J., 2016. The progress test of medicine: the Dutch experience. Perspectives on medical education 5, 51-55.
Turney, B.W., 2007. Anatomy in a modern medical curriculum. Annals of the Royal College of Surgeons of England 89, 104-107.
Verhoeven, B.H., Verwijnen, G.M., Muijtjens, A.M.M., Scherpbier, A.J.J.A., van der Vleuten, C.P.M., 2002. Panel expertise for an Angoff standard setting procedure in progress testing: item writers compared to recently graduated students. Med Educ 36, 860-867.
Verhoeven, B.H., Verwijnen, G.M., Scherpbier, A.J.J.A., Holdrinet, R.S.G., Oeseburg, B., Bulte, J.A., Van der Vleuten, C.P.M., 1998. An analysis of progress test results of PBL and non-PBL students. Med Teach 20, 310-316.
Wass, V., Van der Vleuten, C., Shatzer, J., Jones, R., 2001. Assessment of clinical competence. Lancet 357, 945-949.
Waterston, S.W., Stewart, I.J., 2005. Survey of clinicians' attitudes to the anatomical teaching and knowledge of medical students. Clin Anat 18, 380-384.
Williams, G., Lau, A., 2004. Reform of undergraduate medical teaching in the United Kingdom: a triumph of evangelism over common sense. BMJ 329, 92-94.
Winkelmann, A., 2007. Anatomical dissection as a teaching method in medical school: a review of the evidence. Med Educ 41, 15-22.
Woods, N.N., 2007. Science is fundamental: the role of biomedical knowledge in clinical reasoning. Med Educ 41, 1173-1177.
Woods, N.N., Brooks, L.R., Norman, G.R., 2007a. It all make sense: biomedical knowledge, causal connections and memory in the novice diagnostician. Adv Health Sci Educ 12, 405-415.
Woods, N.N., Brooks, L.R., Norman, G.R., 2007b. The role of biomedical knowledge in diagnosis of difficult clinical cases. Adv Health Sci Educ 12, 417-426.

Woods, N.N., Neville, A.J., Levinson, A.J., Howey, E.H.A., Oczkowski, W.J., Norman, G.R., 2006. The value of basic science in clinical diagnosis. Acad Med 81, S124-S127.
Wrigley, W., van der Vleuten, C.P., Freeman, A., Muijtjens, A., 2012. A systemic framework for the progress test: strengths, constraints and issues: AMEE Guide No. 71. Med Teach 34, 683-697.
Yammine, K., 2014. The current status of anatomy knowledge: where are we now? Where do we need to go and how do we get there? Teaching and learning in medicine 26, 184-188.