Illustrated versus non-illustrated anatomical test items in anatomy course tests and German Medical Licensing examinations (M1)

Olaf Bahlmann[1]

1 Dr. Senckenbergische Anatomie (Institut III), Frankfurt, Germany

Abstract

Illustrated Multiple-choice questions (iMCQs) form an integral part of written tests in anatomy. In iMCQs, the written question refers to various types of figures, e. g. X-ray images, micrographs of histological sections, or drawings of anatomical structures. Since the inclusion of images in MCQs might affect item performance we compared characteristics of anatomical items tested with iMCQs and non-iMCQs in seven tests of anatomy courses and in two written parts of the first section of the German Medical Licensing Examination (M1).

In summary, we compared 25 iMCQs and 163 non-iMCQs from anatomy courses, and 27 iMCQs and 130 non-iMCQs from the written part of the M1 using a nonparametric test for unpaired samples. As a result, there were no significant differences in difficulty and discrimination levels between iMCQs and non-iMCQs, the same applied to an analysis stratified for MCQ formats.

We conclude that the illustrated item format by itself does not seem to affect item difficulty. The present results are consistent with previous retrospective studies which showed no significant differences of test or item characteristics between iMCQs and non-iMCQs.

Keywords: Examination questions, medical illustration, educational measurement, anatomy, anatomy and histology

1. Introduction

Today, there are diverse resources for anatomy assessment. Before the introduction of the Multiple-choice (MC) format in the seventies, state examinations were viva voce [6]. Unstructured oral examinations lack good reliability. Structured oral examinations (SOE) can show good reliability using an item blueprint and a scoring template [13].

There are various formats of test items in written assessment. Currently, the most common is the MC format with four or five answer options. In Single-best answer (SBA) questions, there is only one best (correct) answer. SBA is the most popular MC format. In True/false (T/F) items, all correct answers (more than one) must be marked. Simple T/F items might be acceptable. Multiple T/F items with combinations of answer options, used in medical exams in the past, are no longer recommended [4]. The Extended-matching question (EMQ) includes an option list and at least two item stems, and for each stem, the examinee chooses the single best answer from the list. Multiple-choice examinations show high test reliability [6], [13]. Open questions can be answered by a written essay or keywords (Short-answer question, SAQ). Open questions are more time consuming, compared to MC formats [6]. The Modified essay question (MEQ) is a structured variant of the essay format. In spotter/tag tests, MCQs or SAQs refer to marked (tagged) structures in specimens or images [1], [13].

Multiple-choice questions (MCQs) are widely used in medical exams. In addition, many medical textbooks nowadays include some self-assessment MCQs at the end of a chapter. The National Board of Medical Examiners (NBME) and other authors published guidelines for the creation of MCQs [4], [8]. Visual resources in exam questions should be accurate, complete, relevant and unambiguous [5]. Instructions on how to produce visual material for MCQs and common pitfalls in anatomy MCQs are published [1], [14].

Illustrated MCQs (iMCQs) form an integral part of anatomy tests. Different MCQ formats, e. g. SBA questions or EMQs, can be combined with illustrations. Various illustrations can be included, from x-ray or histological images to photographs of gross preparation specimens or illustrations of functional systems.

An item analysis shows the difficulty and discrimination of individual MCQs. The difficulty index is the proportion of participants choosing the correct answer. Item discrimination is the correlation between the item score and the test score (item total correlation). Good MCQs have a high correlation coefficient [7], [11].

Previous studies did not find significant differences in item or test characteristics between iMCQs and non-iMCQs [3], [9], [12], [15], except for a study on final

year students tested with MCQs presenting a clinical problem. In this study on problem-based radiology questions, illustrated items requiring image interpretation were more difficult compared to questions testing recall of knowledge [10].
However, the integration of illustrations in MCQs might affect item difficulty and overall test difficulty. Therefore, the aim of the present study was to assess characteristics of illustrated and non-illustrated anatomical items from seven anatomy course tests and two written parts of the first section of the German Medical Licensing Examination (M1) in autumn 2015 and 2016.

2. Methods

2.1. Multiple-choice questions

MCQs from seven consecutive anatomy course tests from winter 2014 to summer 2016 provided the basis for this study. First and second year medical and dentistry students participated in the tests. A test with 30 MCQs was written at the end of course one (musculoskeletal system), course two (internal organs), course three (head and neck and neuroanatomy) and the anatomy seminar for medical students. Between 592 and 364 students participated in the anatomy course tests. Medical students of the Goethe-University Frankfurt wrote M1 examinations with 80 anatomy questions each, in autumn 2015 and 2016 with 393 and 330 participants. Anatomy course tests included between 3 and 7 and the written parts of the M1 12 and 15 illustrated anatomical items. Exam papers were evaluated with EvaExam software (Electric paper, Lüneburg, Germany).
MCQs classified as doublets and iMCQs with identical illustrations were excluded from the study. Microsoft Excel was used to calculate item difficulty and discrimination from the raw data. The difficulty index was determined as the mean item score. Item discrimination was calculated as the Pearson product-moment correlation coefficient of the individual item score and the sum score of the remaining items (corrected item discrimination).
Item analyses of M1 questions were produced and are under copyright by the Institute for Medical and Pharmaceutical Exam Questions (IMPP, Mainz, Germany).

2.2. Statistical analysis

Data were inspected and tested for normal distribution (Q-Q plot, Shapiro-Wilkinson test). The Kolmogorov-Smirnov test for unpaired samples was used to compare groups of MCQs. Statistical analysis was performed with GraphPad Prism version 7.00 for Windows (GraphPad Software, La Jolla, California, USA). Data were plotted with median and range. A comparison of iMCQs and non-iMCQs stratified for MCQ formats was performed with the stratified van-Elteren U-test (Bias, Version 11.02, epsilon Verlag, 2016).

3. Results

From anatomy course tests, 25 iMCQs and 163 non-iMCQs were included in this study. IMCQs consisted of 13 histological and 5 radiological images (conventional x-ray or CT), 4 anatomical illustrations, 2 surface anatomy pictures and 1 image of a gross brain section (see Figure 1, translation of the original question). MCQs followed the A-type format (one best answer and four distractors).

In figure No. 1 of the colour supplement the Tractus spinocerebellaris anterior is situated in the area marked by letter:

a) A
b) B
c) C
d) D
e) E

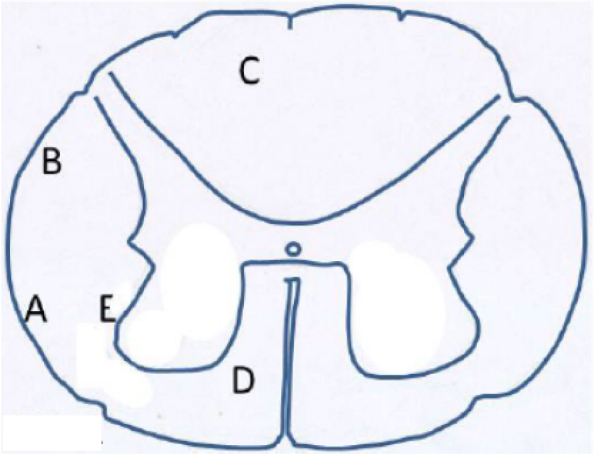

Figure 1: Example of an iMCQ (translation of the original question). The illustrated item format included 13 histological and 5 radiological images, 4 anatomical illustrations, 2 surface anatomy and one brain section image.

Anatomy questions from two M1 examinations with 27 iMCQs and 130 non-iMCQs were also included in this study. 16 images of histological sections, 8 surface anatomy pictures, 1 image of a dissection specimen, 1 image of a body cross section and 1 anatomical illustration were used in iMCQs.
In addition, we stratified items according to MCQ formats. The stratified analysis was run on items with a question in the stem and (short) answer options, positively (Group A) or negatively worded (Group B), and MCQs with statements as answer options, positively (Group C) or negatively worded (Group D). Other formats (sentence completion or matching items) were excluded from the analysis. Median difficulty of iMCQs and non-iMCQs was 0.78 vs 0.76 for anatomy course tests and 0.76 vs 0.82 for the written parts of the M1. The discrimination coefficient was 0.3 vs 0.31 and 0.24 vs 0.315, respectively. As a result, iMCQs and non-iMCQs showed no significant differences in difficulty and discrimination for MCQs of anatomy course tests and written parts of the M1

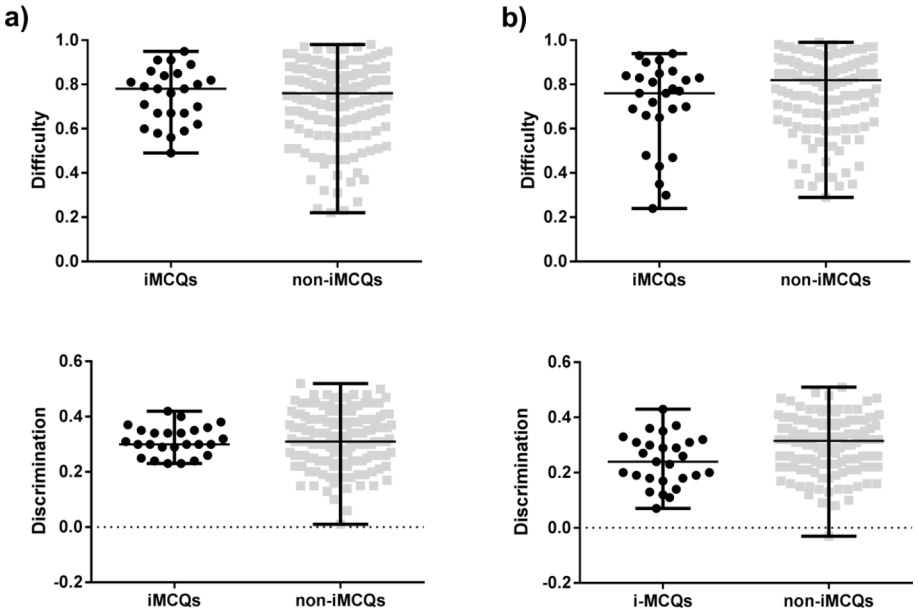

Figure 2: Difficulty and discrimination of iMCQs versus non-iMCQs in anatomy course tests (a) and M1 anatomical items (b). Data shown with median and range. As a result, iMCQs and non-iMCQs were not significantly different in difficulty and discrimination in anatomy course tests and in written parts of the M1.

(p>0.05) (see Figure 2), the same applied to the stratified analysis.

4. Discussion

Visual resources are widely used in anatomy teaching and performance assessment. Each anatomy course test includes some iMCQs, and they are part of the written part of the first section of the Medical Licensing Examination (M1). Thus, in the present study, we were interested in the performance of this item format. Therefore, we compared iMCQs and non-iMCQs in anatomy course tests and the written part of the M1. We found that iMCQs and non-iMCQs did not differ significantly in difficulty and discrimination. The fact that iMCQs and non-iMCQs are based on alternative sources of information, i.e. images and text, does not seem to affect item characteristics.

IMCQs have been assessed previously. Hunt compared two sets of problem-based MCQs in radiology. One set included an image, the other a description of the image, e. g. a radiologist's report. Final year students wrote the sets in two parallel exams. As a result, the set of items with visual content was significantly more difficult. In Hunt's view the results "are consistent with the belief that questions calling for interpretation of data or problem-solving require a higher level of performance or additional skill to that required for questions which supply written descriptions of that data" ([9], p. 420).

In a study on Part 1 FRACS (Fellowship of the Royal Australasian College of Surgeons) exam questions, the authors compared 77 triplets of MCQs in anatomy and pathology. The MCQs presented four answer options. The triplets consisted of a visual and a verbal question of the same content and an additional verbal one of similar content. There were no significant differences in item difficulty and discrimination. The authors argued that their study was limited by a small sample size and that a lower competence in written English of non-native speaker candidates in the FRACS exams might have influenced the results [3].

Vorstenborsch et al. compared 39 EMQs with either an answer list or a labelled anatomical illustration in the item stem. Two test versions were constructed and half of the students wrote each test. Students volunteered for this informal exam, which was similar to the circumstances of an official exam. Using a label, some questions were more and some less difficult, compared to the non-labelled version. Contrary to our study, the authors used extended-matching items instead of MCQs and created closely matched items (labeled image vs answer list). Finally, they were able to compare overall difficulty and reliability of separate test versions. Apart from variable individual effects, the authors did not find overall differences between test versions [15].

Holland et al. reviewed histology exams from three consecutive years with 95 iMCQs and 100 non-iMCQs, and found no significant differences in item difficulty or discrimination [9]. In the present study, we included 25 items of all anatomical subjects including 13 histology questions.

Similarly, in a retrospective analysis on text-only and items with reference images in anatomy examinations, there were no significant differences in difficulty or discrimination between item formats. In this study, the illustration was an addition to the item and did not replace written content, thus images "were considered not to be critical to answering the item" [[12], page 3]. Concerning study design, the studies by Hunt and Vorstenbosch were trial or informal examinations respectively. Students were allocated at random to test groups, and students were not informed about the nature of the examination. Though it

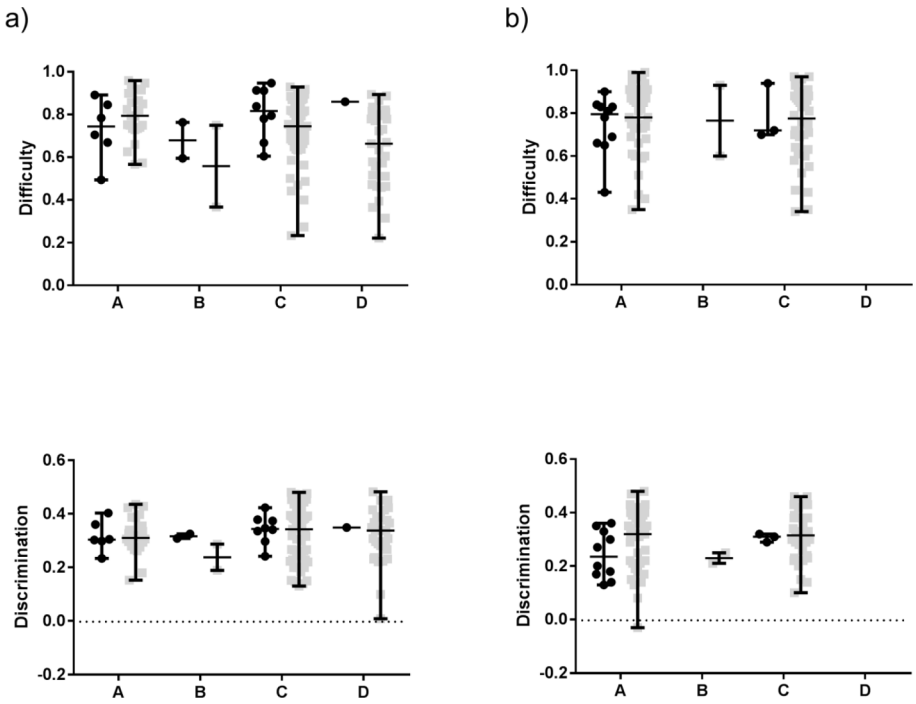

Figure 3: Difficulty and discrimination of iMCQs (●) versus non-iMCQs (□) in anatomy course tests (a) and written parts of the M1 (b), stratified for MCQ formats. Here, MCQs in the (short) answer option format, positively (A) and negatively worded (B), and MCQs in the statement option format, positively (C) and negatively worded (D), are shown, with median and range. As a result, there were no significant differences between iMCQs and non-iMCQs.

was an informal test, test conditions were comparable to an official exam [10]. Each student answered items in both formats [10], [15].

The studies by Buzzard and Hunt included radiological items, which went beyond recall of knowledge and asked for thinking in a clinical context (see item examples) [2], [10]. Hunt categorized items according to clinical setting, supplementary data, interpretation, diagnosis and treatment presented in the question stem and options. In all subgroups, items were more difficult in the illustrated format [10]. In the present study, most of the MCQs cover basic anatomical knowledge on a lower cognitive level.

Hunt showed the increase and decrease of difficulty and discrimination of items created in pairs. 43 out of 70 item pairs increased in difficulty [10]. In the present study, we compared formats of independent items without a pairwise allocation.

In addition, we stratified for MCQ formats (wording and structure of item stems and options) (see Figure 3). However, the integration of illustrations in MCQs had no significant effect on item difficulty and discrimination.

5. Conclusion

In conclusion, iMCQs can be used whenever appropriate. IMCQs can motivate students who are good in visual knowledge and thinking and can be written for lower and higher cognitive levels of exam questions. IMCQs are used to reflect teaching subjects and provide feedback about the effectivity of teaching. Thereby, the introduction of additional visual teaching material can be evaluated by corresponding iMCQs. Using iMCQs, the images must be of sufficient quality and size and accurately labeled. According to constructive alignment, a test blueprint helps choosing iMCQs for the exam. Different kinds of illustrations (histological images, x-rays) will reflect the diversity of visual input in medicine. Checking the quality of iMCQs will also improve students' learning from trial exam questions. Finally, the results of this study might reassure question writers to use iMCQs.

Acknowledgements

The author would like to acknowledge Professor Eva Herrmann for statistical advice, Professor Jörg Stehle and Professor Frank Nürnberger for helpful comments and the latter for the iMCQ example.

Competing interests

The author declares that he has no competing interests.

References

1. Brenner E, Chirculescu AR, Reblet C, Smith C. Assessment in anatomy. Eur J Anat. 2015;19(1):105-124.
2. Buzzard AJ, Bandaranayake R, Harvey C. How to produce visual material for multiple choice examinations. Med Teach. 1987;9(4):451-456. DOI: 10.3109/01421598709008341
3. Buzzard AJ, Bandaranayake RC. Comparison of the performance of visual and verbal multiple-choice questions. Aus N Z J Surg. 1991;61(8):614-618. DOI: 10.1111/j.1445-2197.1991.tb00302.x

4. Case SM, Swanson DB. National Board of Medical Examiners. Constructing written test questions for the basic and clinical sciences, 3rd ed. Philadelphia: National Board of Medical Examiners; 2001. Zugänglich unter/available from: http://www.nbme.org/pdf/itemwriting_2003/2003iwgwhole.pdf

5. Crisp V, Sweiry E. Can a picture ruin a thousand words? The effect of visual resources in exam questions. Educ Res. 2006;48(2):139-154. DOI: 10.1080/00131880600732249

6. Elmer A, Grifka J. Vergleich von Prüfungsmethoden in der medizinischen Ausbildung. Gesundheitswesen (Suppl Med Ausbild). 1998;15(Suppl1):14-17.

7. Haladyna TM. Developing and validating multiple-choice test items, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates; 2004.

8. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ. 2002;15(3):309-334. DOI: 10.1207/S15324818AME1503_5

9. Holland J, O'Sullivan R, Arnett R. Is a picture worth a thousand words: an analysis of the difficulty and discrimination parameters of illustrated vs. text-alone vignettes in histology multiple choice questions. BMC Med Educ. 2015;15:184. Doc184. DOI: 10.1186/s12909-015-0452-9

10. Hunt DR. Illustrated multiple choice examinations. Med Educ. 1978;12(6):417-420. DOI: 10.1111/j.1365-2923.1978.tb01420.x

11. Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. Basic quantitative analyses of medical examinations. GMS Z Med Ausbild. 2006;23(3):Doc53. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2006-23/zma000272.shtml

12. Notebaert AJ. The effect of images on item statistics in multiple choice anatomy examinations. Anat Sci Educ. 2017;10(1):68-78. DOI: 10.1002/ase.1637

13. Samarasekera DD, Gopalakirshnakone P, Gwee MC. Assessing anatomy as a basic medical science. In: Chan LK, Pawlina W (Hrsg). Teaching anatomy: a practical guide. Bern: Springer International Publishing; 2015. S.279-289. DOI: 10.1007/978-3-319-08930-0_31

14. Vahalia KV, Subramaniam K, Marks SC Jr, De Souza EJ. The use of multiple-choice tests in anatomy: Common pitfalls and how to avoid them. Clin Anat. 1995;8(1):61-65. DOI: 10.1002/ca.980080111

15. Vorstenbosch MA, Klaassen TP, Kooloos, JG, Bolhuis SM, Laan RF. Do images influence assessment in anatomy? Exploring the effect of images on item difficulty and item discrimination. Anat Sci Educ. 2013;6(1):29-41. DOI: 10.1002/ase.1290

Corresponding author:
Dr. med. Olaf Bahlmann
Dr. Senckenbergische Anatomie (Institut III), Theodor-Stern-Kai 7, D-60590 Frankfurt, Germany, Phone: +49 (0)69/6301-6046, Fax: +49 (0)69/6301-6902
bahlmann@med.uni-frankfurt.de

Please cite as
Bahlmann O. Illustrated versus non-illustrated anatomical test items in anatomy course tests and German Medical Licensing examinations (M1). GMS J Med Educ. 2018;35(2):Doc25. DOI: 10.3205/zma001172, URN: urn:nbn:de:0183-zma0011725

This article is freely available from
http://www.egms.de/en/journals/zma/2018-35/zma001172.shtml

Received: *2017-05-31*
Revised: *2018-02-02*
Accepted: *2018-03-04*
Published: *2018-05-15*

Bebilderte versus nicht bebilderte anatomische Testitems in Anatomiekursklausuren und Medizinischen Staatsexamensprüfungen (M1)

Olaf Bahlmann[1]

1 Dr. Senckenbergische Anatomie (Institut III), Frankfurt, Deutschland

Zusammenfassung

Bebilderte Multiple-Choice- (MC) Fragen sind ein integraler Bestandteil von schriftlichen Prüfungen in der Anatomie. In bebilderten MC-Fragen bezieht sich die schriftliche Frage auf verschiedene Typen von Abbildungen wie Röntgenaufnahmen, Mikrofotografien von histologischen Schnitten oder Zeichnungen von anatomischen Strukturen. Da das Hereinnehmen von Abbildungen in MC-Fragen das Abschneiden der Items beeinflussen kann, verglichen wir die Charakteristika von anatomischen Items getestet mit bebilderten und nicht bebilderten MC-Fragen in sieben Anatomieklausuren und in zwei schriftlichen Teilen des Ersten Abschnitts der Ärztlichen Prüfung (M1).

In dieser Studie verglichen wir 25 bebilderte und 163 nicht bebilderte MC-Fragen aus Anatomieklausuren und 27 bebilderte und 130 nicht bebilderte MC-Fragen aus dem schriftlichen Teil des M1 mit einem nicht parametrischen Test für ungepaarte Stichproben. Als Ergebnis waren keine signifikanten Unterschiede im Schwierigkeits- und Trennschärfeniveau zwischen bebilderten und nicht bebilderten MC-Fragen vorhanden, dasselbe ergab sich in einer nach MC-Frageformaten stratifizierten Analyse.

Wir schließen daraus, dass das bebilderte Itemformat für sich die Itemschwierigkeit nicht zu beeinflussen scheint. Die aktuellen Ergebnisse stimmen mit früheren retrospektiven Studien überein, die keine signifikanten Unterschiede zwischen Test- und Itemcharakteristika zwischen bebilderten und nicht bebilderten MC-Fragen zeigten.

Schlüsselwörter: Prüfungsfragen, Medizinische Abbildung, Lernerfolgsmessung, Anatomie, Anatomie und Histologie

1. Einleitung

Heutzutage existieren verschiedene Ressourcen für die Lernerfolgsmessung in Anatomie. Vor der Einführung des Multiple-Choice- (MC) Formats in den Siebzigerjahren waren Staatsexamina viva voce [6]. Unstrukturierten mündlichen Prüfungen fehlt eine gute Reliabilität. Strukturierte mündliche Prüfungen können unter Verwendung eines Blueprints und einer Bewertungsschablone eine gute Reliabilität aufweisen [13].

In der schriftlichen Lernerfolgsmessung gibt es verschiedene Formate von Testitems. Derzeit ist das MC-Format mit vier oder fünf Antwortoptionen das häufigste. In Single-Best-Answer-Fragen gibt es nur eine richtige Antwort. Single-Best-Answer-Fragen sind das beliebteste MC-Format. In Richtig/Falsch-Items müssen alle richtigen Antworten (mehr als eine) markiert werden. Einfache Richtig/Falsch-Items mögen akzeptabel sein, Richtig/Falsch-Items mit Antwortkombinationen, wie sie in medizinischen Examina in der Vergangenheit verwendet wurden, werden nicht mehr empfohlen [4]. Die Extended-Matching-Frage beinhaltet eine Liste mit Antwortoptionen und mindestens zwei Fragestämme und der Prüfling muss für jeden Fragestamm die richtige Antwort aus der Liste auswählen. MC-Prüfungen zeigen eine gute Reliabilität [6], [13]. Offene Fragen können durch einen schriftlichen Essay oder Stichworte beantwortet werden (Short-answer-Frage). Offene Fragen sind, verglichen mit MC-Formaten, zeitaufwendiger [6]. Die modifizierte Essayfrage ist eine strukturierte Variante des Essayformats. In sogenannten Spotter/Tag-Tests beziehen sich MC- oder Short-answer-Fragen auf in Abbildungen oder Präparaten markierte Strukturen [1], [13].

MC-Fragen werden häufig in medizinischen Examina verwendet. Außerdem beinhalten viele Medizinlehrbücher MC-Fragen zur Selbstüberprüfung am Ende eines Kapitels. Das National Board of Medical Examiners (NBME) und andere Autoren haben Leitlinien für die Erstellung von MC-Fragen veröffentlicht [4], [8]. Visuelle Ressourcen in Prüfungsfragen sollten akkurat, komplett, relevant und unmissverständlich sein [5]. Wie visuelles Material für

MC-Fragen erstellt wird und häufige Fallstricke in anatomischen MC-Fragen wurden publiziert [1], [14].
Bebilderte MC-Fragen sind ein integraler Bestandteil von Anatomieklausuren. Verschiedene MC-Frageformate, z.B. Single-Best-Answer- oder Extended-Matching-Fragen, können mit Abbildungen kombiniert werden. Verschiedene Abbildungen können hinzugenommen werden, von radiologischen oder histologischen Abbildungen bis zu Fotografien von makroskopischen Präparaten oder Abbildungen funktioneller Systeme.
Eine Itemanalyse zeigt die Schwierigkeit und Trennschärfe individueller MC-Fragen. Der Schwierigkeitsindex ist der Anteil an Teilnehmern, die das Item richtig beantwortet haben. Die Trennschärfe ist die Korrelation zwischen dem Item- und Testergebnis (Item-Gesamt-Korrelation). Gute MC-Fragen haben einen hohen Korrelationskoeffizienten [7], [11].
Frühere Studien fanden keine signifikanten Unterschiede von Item- oder Testcharakteristika zwischen bebilderten und nicht bebilderten MC-Fragen [3], [9], [12], [15]. Ausnahme war eine Studie an Studierenden im letzten Studienjahr, die mit MC-Fragen mit einer klinischen Problemstellung getestet wurden. In dieser Studie mit problembasierten Radiologiefragen waren bebilderte Items, die eine Bildinterpretation erforderten, schwieriger als Fragen, die das Erinnern von Wissensinhalten prüften [10].
Die Hereinnahme von Abbildungen in MC-Fragen beeinflusst möglicherweise die Itemschwierigkeit und damit die Schwierigkeit der Prüfung. Daher war das Ziel der vorliegenden Arbeit, die Charakteristika von bebilderten und nicht bebilderten anatomischen Items aus sieben Anatomiekursklausuren und aus zwei schriftlichen Teilen des Ersten Abschnitts der Ärztlichen Prüfung (M1) aus dem Herbst 2015 und 2016 zu untersuchen.

2. Methoden

2.1. Multiple-Choice-Fragen

MC-Fragen aus sieben aufeinanderfolgenden Anatomiekursklausuren von Winter 2014 bis Sommer 2016 bildeten die Grundlage für diese Studie. Human- und Zahnmedizinstudierende aus dem ersten und zweiten Studienjahr nahmen an den Klausuren teil. Eine Klausur mit 30 Fragen wurde am Ende des ersten Kurses (Muskuloskelettales System), zweiten Kurses (Innere Organe), dritten Kurses (Kopf- und Hals- und Neuroanatomie) und dem Anatomieseminar für Humanmedizinstudierende geschrieben. Zwischen 592 und 364 Studierende nahmen an den Anatomiekursklausuren teil. Medizinstudierende der Goethe-Universität Frankfurt schrieben die M1-Prüfungen mit je 80 Anatomiefragen im Herbst 2015 und 2016 mit 393 und 330 Teilnehmern. Die Anatomiekursklausuren beinhalteten zwischen 3 und 7 und die schriftlichen Teile des M1 12 und 15 bebilderte anatomische Items. Die Klausurbögen wurden mit der EvaExam Software (Electric paper, Lüneburg, Deutschland) evaluiert.
MC-Fragen, die als Doppel klassifiziert wurden, und bebilderte MC-Fragen mit identischen Abbildungen wurden von der Studie ausgeschlossen. Microsoft Excel wurde für die Berechnung der Itemschwierigkeit und –trennschärfe aus den Rohdaten verwendet. Der Schwierigkeitsindex wurde als mittleres Itemergebnis bestimmt. Die Itemtrennschärfe wurde als der Pearson-Produkt-Moment-Korrelationskoeffizient des individuellen Itemergebnisses und des Summenergebnisses der restlichen Items (korrigierte Itemtrennschärfe) berechnet.
Die Itemanalysen der M1-Fragen wurden vom Institut für Medizinische und Pharmazeutische Prüfungsfragen (IMPP, Mainz, Deutschland) erstellt und sind urheberrechtlich geschützt.

2.2. Statistische Auswertung

Die Daten wurden auf Normalverteilung inspiziert und getestet (Q-Q-Graph, Shapiro-Wilkinson-Test). Der Kolmogorov-Smirnov-Test für ungepaarte Stichproben wurde für den Vergleich der MC-Fragegruppen verwendet. Die statistische Auswertung wurde mit GraphPad Prism Version 7.00 für Windows (GraphPad Software, La Jolla, Kalifornien, USA) durchgeführt. Die Daten wurden mit Median und Spannweite aufgetragen. Ein Vergleich von bebilderten und nicht bebilderten MC-Fragen, stratifiziert nach MC-Frageformaten, wurde mit dem stratifizierten van-Elteren-U-Test (Bias, Version 11.02, epsilon-Verlag, 2016) vorgenommen.

3. Ergebnisse

Aus den Anatomiekursklausuren wurden 25 bebilderte und 163 nicht bebilderte MC-Fragen in diese Studie aufgenommen. Bebilderte MC-Fragen umfassten 13 histologische und 5 radiologische Abbildungen (konventionelle Röntgenaufnahmen oder CT), 4 anatomische (Schema)-Zeichnungen, 2 Anatomie-in-vivo Abbildungen und eine Abbildung einer Hirnscheibe (siehe Abbildung 1). Die MC-Fragen beinhalteten eine am wahrscheinlichsten zutreffende/nicht zutreffende Antwort und vier Distraktoren (A-Fragenformattyp).
Anatomiefragen aus zwei M1-Prüfungen mit 27 bebilderten und 130 nicht bebilderten MC-Fragen wurden zudem in diese Studie aufgenommen. 16 histologische und 8 Anatomie-in-vivo Abbildungen, eine Abbildung eines anatomischen Präparats, eine Abbildung einer Körperscheibe und eine anatomische Schema-Zeichnung wurden in den bebilderten MC-Fragen verwendet.
Außerdem stratifizierten wir die Items nach MC-Frageformaten. Die stratifizierte Auswertung wurde an Items mit einer Fragestellung im Stamm und (kurzen) Antwortoptionen, positiv (Gruppe A) oder negativ formuliert (Gruppe B), und MC-Fragen mit Aussagen als Antwortoptionen, positiv (Gruppe C) oder negativ formuliert (Gruppe D), vorgenommen. Andere Formate (Satzvervollständigung oder Zuordnungsitems) wurden von der Auswertung ausgeschlossen.

In der Abbildung 1 der Bildbeilage befindet sich der Tractus spinocerebellaris anterior im Bereich der Markierung mit dem Buchstaben:

a) A
b) B
c) C
d) D
e) E

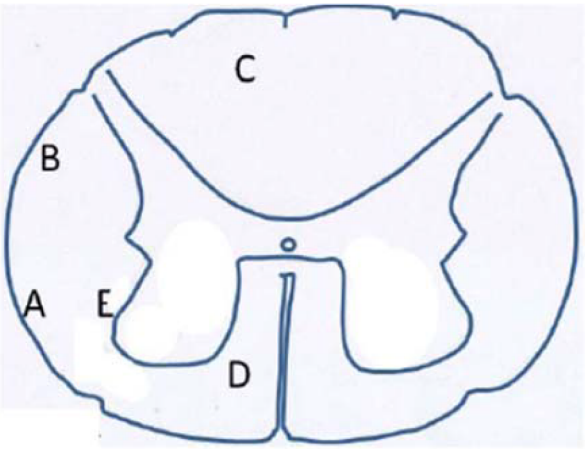

Abbildung 1: Beispiel einer bebilderten MC-Frage. Das bebilderte Itemformat beinhaltete 13 histologische und 5 radiologische Abbildungen, 4 anatomische Zeichnungen, 2 Anatomie in vivo Abbildungen und eine Abbildung einer Hirnscheibe.

Die mittlere Schwierigkeit von bebilderten und nicht bebilderten MC-Fragen war 0.78 vs. 0.76 in Anatomiekursklausuren und 0.76 vs. 0.82 in den schriftlichen Teilen des M1. Der Trennschärfekoeffizient war entsprechend 0.3 vs. 0.31 und 0.24 vs. 0.315. Als Ergebnis zeigten bebilderte und nicht bebilderte MC-Fragen keine signifikanten Unterschiede in Schwierigkeit und Trennschärfe in Anatomiekursklausuren und den schriftlichen Teilen des M1 ($p>0.05$) (siehe Abbildung 2), was ebenfalls für die stratifizierte Auswertung zutraf.

4. Diskussion

Visuelle Ressourcen werden häufig im Anatomieunterricht und der Leistungsüberprüfung gebraucht. Jede Anatomieklausur beinhaltet bebilderte MC-Fragen, und sie sind Bestandteil des schriftlichen Teils des M1. Daher interessierten wir uns in der vorliegenden Studie für das Abschneiden dieses Itemformats. Dazu verglichen wir bebilderte und nicht bebilderte MC-Fragen in Anatomiekursklausuren und dem schriftlichen Teil des M1. Wir fanden, dass sich bebilderte und nicht bebilderte MC-Fragen nicht signifikant in der Schwierigkeit und Trennschärfe unterschieden. Dass bebilderte und nicht bebilderte MC-Fragen auf verschiedenen Ressourcen basieren, d. h. Abbildungen und Text, scheint die Itemcharakteristika nicht zu beeinflussen.

Bebilderte MC-Fragen wurden in der Vergangenheit bereits untersucht. Hunt verglich zwei Sets an problembasierten MC-Fragen in der Radiologie. Ein Set beinhaltete eine Abbildung, das andere eine Beschreibung der Abbildung, z. B. einen Radiologiebericht. Studierende im letzten Studienjahr schrieben die Sets in zwei parallelen Prüfungen. Als Ergebnis war das Set mit visuellem Inhalt signifikant schwieriger. Nach Auffassung von Hunt waren die Ergebnisse "übereinstimmend mit der Ansicht, dass Fragen, die eine Interpretation der Daten oder eine Problemlösung verlangen, ein höheres Leistungslevel oder zusätzliche Fähigkeiten erfordern als Fragen, die eine schriftliche Beschreibung der Daten bieten" ([10], S. 420).

In einer Studie an Fragen aus dem ersten Teil des FRACS-Examens (Fellowship of the Royal Australasian College of Surgeons) verglichen die Autoren 77 MC-Fragentripletts zur Anatomie und Pathologie. Die MC-Fragen boten vier Antwortoptionen. Die Tripletts bestanden aus einer visuellen und verbalen Frage desselben und einer zusätzlichen verbalen Frage vergleichbaren Inhalts. Es ergaben sich keine signifikanten Unterschiede in Itemschwierigkeit und -trennschärfe. Die Autoren argumentierten, dass ihre Studie durch eine geringe Fallzahl limitiert sei und dass eine niedrigere Kompetenz in der englischen Schriftsprache von Nicht-Muttersprachlern die Ergebnisse des FRACS-Examens beeinflusst haben könnte [3].

Vorstenborsch et al. verglichen 39 Extended-Matching-Fragen mit entweder einer Antwortliste oder einer beschrifteten anatomischen Abbildung im Fragestamm. Es wurden zwei Testversionen erstellt und die Hälfte der Studierenden schrieben beide Tests. Die Studierenden nahmen freiwillig an der informellen Prüfung teil, die vergleichbar den Bedingungen einer offiziellen Prüfung war. Manche der beschrifteten Fragen waren schwieriger und andere dagegen weniger schwierig, verglichen mit der nicht beschrifteten Version. Anders als in unserer Studie verwendeten die Autoren Extended-Matching- anstatt MC-Fragen und erstellten eng abgestimmte Items (beschriftete Abbildungen bzw. Antwortliste). Schlussendlich war es ihnen möglich, die Gesamtschwierigkeit und Reliabilität der separaten Testversionen zu vergleichen. Abgesehen von variablen Einzeleffekten fanden die Autoren keine Gesamtunterschiede zwischen den Testversionen [15].

Holland et al. untersuchten Histologieklausuren dreier aufeinanderfolgender Jahre mit 95 bebilderten und 100 nicht bebilderten MC-Fragen und fanden keine signifikanten Unterschiede in Itemschwierigkeit oder –trennschärfe [9]. In die vorliegende Studie nahmen wir 25 Items aller anatomischer Gebiete auf inklusive 13 histologische Fragen.

In ähnlicher Weise waren in einer retrospektiven Auswertung von Textaufgaben und Items mit Referenzabbildungen aus Anatomieprüfungen keine signifikanten Unterschiede in Schwierigkeit oder Trennschärfe zwischen den Itemformaten vorhanden. In dieser Studie ergänzten die Abbildungen das Item und ersetzten nicht den schriftlichen Inhalt, somit "waren [die Abbildungen] nicht als kritisch für die Beantwortung des Items anzusehen" ([12], Seite 3). Was das Studiendesign betrifft, waren die Studien von Hunt und Vorstenbosch Probeklausuren bzw. informelle Prüfungen. Die Studierenden wurden randomi-

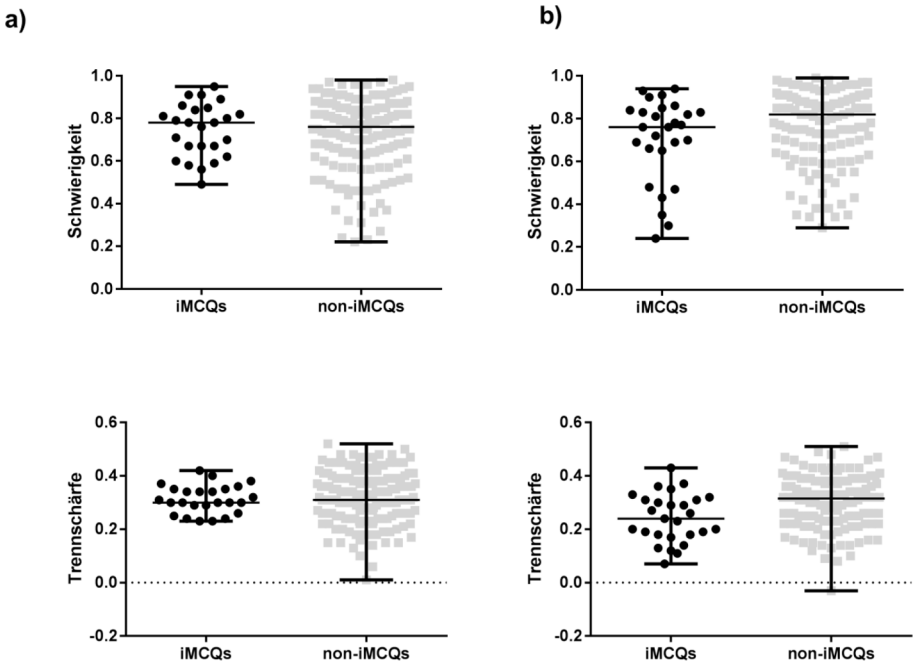

Abbildung 2: Schwierigkeit und Trennschärfe bebilderter (IMCQs) versus nicht bebilderter MC-Fragen (non-IMCQs) in Anatomiekursklausuren (a) und M1-Anatomieitems (b). Daten aufgetragen mit Median und Spannweite. Bebilderte und nicht bebilderte MC-Fragen unterschieden sich nicht signifikant in ihrer Schwierigkeit und Trennschärfe in Anatomiekursklausuren und in den schriftlichen Teilen des M1.

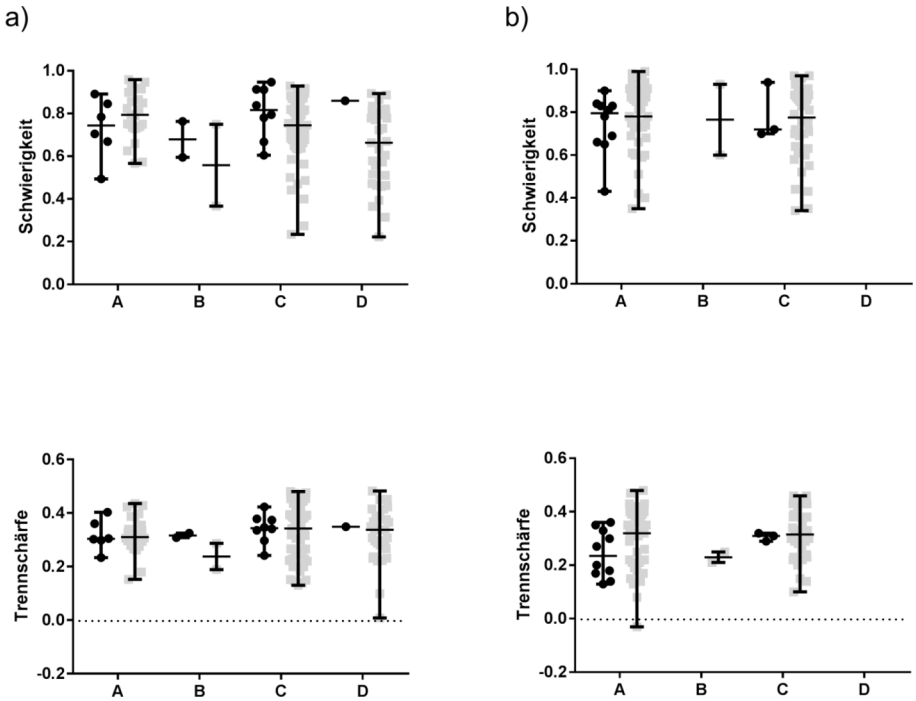

Abbildung 3: Schwierigkeit und Trennschärfe bebilderter (IMCQs) (●) versus nicht bebilderter MC-Fragen (non-IMCQs) (□) in Anatomiekursklausuren (a) und dem schriftlichen Teil des M1 (b), stratifiziert nach MC-Frageformaten. Dargestellt sind MC-Fragen im (kurzen) Antwortoptionsformat, positiv (A) und negativ formuliert (B), und MC-Fragen im Aussageoptionsformat, positiv (C) und negativ (D) formuliert, mit Median und Spannweite. Es ergaben sich keine signifikanten Unterschiede zwischen bebilderten und nicht bebilderten MC-Fragen.

siert den Testgruppen zugewiesen und nicht über die Testform unterrichtet. Obwohl es sich um eine informelle Prüfung handelte, waren die Testbedingungen vergleichbar einer offiziellen Prüfung [10]. Jede(r) Studierende beantwortete die Items in beiden Formaten [10], [15].

Die Studien von Buzzard und Hunt beinhalteten radiologische Items, die über eine Erinnerung von Wissensinhalten hinaus gingen und Denken im klinischen Kontext verlangten (vgl. Itembeispiele) [2], [10]. Hunt kategorisierte die Items nach klinischem Szenario, supplementären Daten, Interpretation, Diagnose und Behandlung im Fra-

gestamm und Optionen. Die Items waren im bebilderten Format in allen Untergruppen schwieriger [10]. In der vorliegenden Studie deckten die meisten Items grundlegendes anatomisches Wissen auf niedrigerem kognitivem Niveau ab.
Hunt zeigte den Anstieg und die Abnahme der Schwierigkeit und Trennschärfe an Itempaaren. In 43 von 70 Itempaaren nahm die Schwierigkeit zu [10]. In der vorliegenden Studie verglichen wir die Formate von unabhängigen Items und nahmen keine paarweise Zuordnung vor. Außerdem stratifizierten wir nach MC-Frageformaten (Formulierung und Struktur des Fragestamms und der Optionen) (siehe Abbildung 3). Die Hereinnahme von Abbildungen in MC-Fragen hatte keinen signifikanten Effekt auf die Itemschwierigkeit und -trennschärfe.

5. Schlussfolgerung

Bebilderte MC-Fragen sind immer, wenn es geeignet erscheint, einsetzbar. Bebilderte MC-Fragen können Studierende motivieren, die gut sind in visuellem Wissen und Denken, und sie können für niedrigere und höhere kognitive Niveaus geschrieben werden. Bebilderte MC-Fragen werden genutzt, um Lehrgegenstände zu reflektieren und Rückmeldung über die Effektivität des Lehrens zu erhalten. Dadurch kann die Einführung von zusätzlichen visuellen Lehrgegenständen durch entsprechende bebilderte MC-Fragen evaluiert werden. Bei der Verwendung von bebilderten MC-Fragen ist darauf zu achten, dass die Abbildungen von ausreichender Qualität und Größe und sorgfältig beschriftet sind. Gemäß einem Constructive Alignment hilft ein Blueprint bei der Auswahl von bebilderten MC-Fragen für die Klausur. Verschiedenartige Abbildungen (histologische Abbildungen, Röntgenaufnahmen) reflektieren die Vielseitigkeit der bildlichen Informationen in der Medizin. Die Überprüfung der Qualität von bebilderten MC-Fragen wird auch das studentische Lernen an Probeklausurfragen verbessern. Schlussendlich werden die Ergebnisse der vorliegenden Studie vielleicht Frageschreiber bestärken, bebilderte MC-Fragen zu verwenden.

Danksagung

Der Autor möchte Frau Professor Eva Herrmann für die statistische Beratung, Herrn Professor Jörg Stehle und Herrn Professor Frank Nürnberger für hilfreiche Kommentare und dem Letztgenannten für das Beispiel einer bebilderten MC Frage danken.

Interessenkonflikt

Der Autor erklärt, dass er keine Interessenkonflikte im Zusammenhang mit diesem Artikel hat.

Literatur

1. Brenner E, Chirculescu AR, Reblet C, Smith C. Assessment in anatomy. Eur J Anat. 2015;19(1):105-124.
2. Buzzard AJ, Bandaranayake R, Harvey C. How to produce visual material for multiple choice examinations. Med Teach. 1987;9(4):451-456. DOI: 10.3109/01421598709008341
3. Buzzard AJ, Bandaranayake RC. Comparison of the performance of visual and verbal multiple-choice questions. Aus N Z J Surg. 1991;61(8):614-618. DOI: 10.1111/j.1445-2197.1991.tb00302.x
4. Case SM, Swanson DB. National Board of Medical Examiners. Constructing written test questions for the basic and clinical sciences, 3rd ed. Philadelphia: National Board of Medical Examiners; 2001. Zugänglich unter/available from: http://www.nbme.org/pdf/itemwriting_2003/2003iwgwhole.pdf
5. Crisp V, Sweiry E. Can a picture ruin a thousand words? The effect of visual resources in exam questions. Educ Res. 2006;48(2):139-154. DOI: 10.1080/00131880600732249
6. Elmer A, Grifka J. Vergleich von Prüfungsmethoden in der medizinischen Ausbildung. Gesundheitswesen (Suppl Med Ausbild). 1998;15(Suppl1):14-17.
7. Haladyna TM. Developing and validating multiple-choice test items, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates; 2004.
8. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ. 2002;15(3):309-334. DOI: 10.1207/S15324818AME1503_5
9. Holland J, O'Sullivan R, Arnett R. Is a picture worth a thousand words: an analysis of the difficulty and discrimination parameters of illustrated vs. text-alone vignettes in histology multiple choice questions. BMC Med Educ. 2015;15:184. Doc184. DOI: 10.1186/s12909-015-0452-9
10. Hunt DR. Illustrated multiple choice examinations. Med Educ. 1978;12(6):417-420. DOI: 10.1111/j.1365-2923.1978.tb01420.x
11. Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. Basic quantitative analyses of medical examinations. GMS Z Med Ausbild. 2006;23(3):Doc53. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2006-23/zma000272.shtml
12. Notebaert AJ. The effect of images on item statistics in multiple choice anatomy examinations. Anat Sci Educ. 2017;10(1):68-78. DOI: 10.1002/ase.1637
13. Samarasekera DD, Gopalakirshnakone P, Gwee MC. Assessing anatomy as a basic medical science. In: Chan LK, Pawlina W (Hrsg). Teaching anatomy: a practical guide. Bern: Springer International Publishing; 2015. S.279-289. DOI: 10.1007/978-3-319-08930-0_31
14. Vahalia KV, Subramaniam K, Marks SC Jr, De Souza EJ. The use of multiple-choice tests in anatomy: Common pitfalls and how to avoid them. Clin Anat. 1995;8(1):61-65. DOI: 10.1002/ca.980080111
15. Vorstenbosch MA, Klaassen TP, Kooloos, JG, Bolhuis SM, Laan RF. Do images influence assessment in anatomy? Exploring the effect of images on item difficulty and item discrimination. Anat Sci Educ. 2013;6(1):29-41. DOI: 10.1002/ase.1290

Korrespondenzadresse:
Dr. med. Olaf Bahlmann
Dr. Senckenbergische Anatomie (Institut III),
Theodor-Stern-Kai 7, 60590 Frankfurt, Deutschland, Tel.:
+49 (0)69/6301-6046, Fax: +49 (0)69/6301-6902
bahlmann@med.uni-frankfurt.de

Bitte zitieren als
Bahlmann O. Illustrated versus non-illustrated anatomical test items in anatomy course tests and German Medical Licensing examinations (M1). GMS J Med Educ. 2018;35(2):Doc25.
DOI: 10.3205/zma001172, URN: urn:nbn:de:0183-zma0011725

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/zma/2018-35/zma001172.shtml

Eingereicht: *31.05.2017*
Überarbeitet: *02.02.2018*
Angenommen: *04.03.2018*
Veröffentlicht: *15.05.2018*