Assessment of anatomical knowledge by practical examinations: the effect of question design on student performance

Article (Accepted Version)

Assessment of Anatomical Knowledge by Practical Examinations: The effect of Question Design on Student Performance

Mandeep Gill Sagoo[1], Claire France Smith[2], Edward Gosden[3]

[1]Department of Anatomy and Human Sciences, King's College, London, Guy's campus, Great Maze Pond, London, SE19RT, United Kingdom. T: +44 2078486064, E-mail: mgsagoo@gmail.com, mandeep.gill@kcl.ac.uk

[2]Department of Anatomy, Brighton and Sussex Medical School, Biology Road, University of Sussex, Falmer, Brighton, BN19PX, United Kingdom. T: (0123) 877890, F: (01273) 877576, E-mail: C.Smith@bsms.ac.uk

[3]Research Methods Group, Institute of Health and Biomedical Innovation, 60 Musk Avenue, Kelvin Grove Queensland University of Technology, QLD 4059, Australia. T: +61 7313861981, E-mail: edward.gosden@qut.edu.au

Running Title: Anatomical Question Design and Students Performance

Correspondence to: Mrs. Mandeep Gill Sagoo, Department of Anatomy and Human Sciences, King's College, London, Guy's campus, Great Maze Pond, London, SE19RT, United Kingdom, T: +44 2078486064, M: +447976221172, E-mail: mgsagoo@gmail.com, mandeep.gill@kcl.ac.uk

ABSTRACT

The Objective Structural Practical Examination (OSPE) is a timed examination that assesses topographical and/or applied knowledge of anatomy with the use of cadaveric resources and medical images. This study investigated whether elements of question design (provision of clinical context, type of visual resources used, gender context and difficulty) of an anatomy question affected students' performance and also whether there was any effect of basic demography or participation in various voluntary activities. Study participants were second year medical students (n = 150), 83 of whom consented to fill in a questionnaire collecting demographics, revision preferences and assessment preferences. The examination scores were matched with students' responses collected on the questionnaire and all data analyzed by multiple linear regression. Difficulty of the question was the only design element found to be significantly associated with the number of students that answered correctly (p=. 001); clinical context, visual resources used and gender of the question were not significant. When individual students marks were analyzed along with the questionnaire data, only the students' interest in participating in department's demonstrator program was a significant predictor of a high individual score, gender of the students showed a strong trend towards significance, with female students scoring on average higher than male students. The two part OSPE questions were dissociated and analyzed using binary logistic regression to determine whether a correct answer to part 1 was predictive of a correct answer to part 2, but no association was found.

Keywords: Anatomical sciences/medical education; gross anatomy education; undergraduate education; anatomy assessments; practical examinations; OSPE; anatomical visual resources; clinical context; medical students' scores; medical students' views.

INTRODUCTION

The importance of anatomy in the field of medicine and its impact on medical students' clinical careers is well known (Smith and Mathias 2011). However, curriculum design, teaching and assessment methods are still the subject of considerable debate. Recent thinking favors approaches that facilitate the application of knowledge in practice (McHanwell et al., 2007), but both teaching and assessment practices come under considerable pressure from competing space and time demands (Moxham et al. 2011). A variety of assessment methods have been developed over the course of time, most attempt to combine the ability to identify gaps in students' basic science knowledge with assessing their ability to apply that knowledge, these methods group into oral examinations (viva), written examinations (paper or online) and practical examinations (Rowland et al., 2011). Oral examinations are now seldom employed in the UK because of perceptions of bias and low reliability, and also because they are time-consuming (Smith and Mcmanus, 2015) but these have been widely used in the US (Clough and Lehr, 1996; Fabrizio, 2013). Written examinations are common and range from Single Best Answer (SBA) multiple-choice questions, through a variety of techniques ranging from spotter questions and steeplechases to essays (Schubert et al., 2009; Inuwa et al., 2011, 2012; Yaquinddin et al., 2013). The pressure of marking forces examinations is

towards SBAs because of the ease and the possibility of electronic marking (Daly, 2010).

This study is concerned with the practical examination of anatomy, particularly the Objective Structural Practical Examination (OSPE) which is widely accepted because of its perceived ability to assess a range of theoretical, applied and procedural skills simultaneously (Menezes et al., 2011 and Nayar et al., 1986). A classical anatomical OSPE involves students moving around a series of stations, each of which may contain specimens, prosected cadavers, models or images. Questions are typically in two parts, the first requiring identification of a structure and the second linking the structure to its function. Time allowed at each station is often limited (1-1.5 min), and students have to move to the next station at an audible signal. OSPEs differ from standard spotter examinations, in that they seek to go beyond the simple identification of a structure and so do not simply test recall of structural knowledge. Recent developments in OSPE methodology have moved away from dissection-room scenarios (because of a lack of accessibility and/or availability of cadaveric resources) into online OSPE-like assessments (Gunderman, 2008; Krippendorf et al., 2008; Inuwa et al., 2011, 2012), which are more practical and less labor intensive. However, studies by Fitzgerald et al. (2009) and Smith and Mathias (2011) have emphasized the importance of the three dimensional experience given by cadaveric specimens, and highlighted the necessity for anatomy to be taught and assessed in this context. Studies by Regan de Bere and Mattick (2010) and Vorstenbosch et al. (2015) suggested that junior doctors actively used anatomical knowledge during a consultation with a patient, and that this use relied

on adequate visual representations in memory. (Schnotz and Baadte, 2015). However, a number of studies have shown that the use of a variety of different visual resources does not have a significant effect on students' performance (Shubert et al., 2009; Inuwa et al., 2011, 2012).

This study focuses on investigating the relationship between the design of anatomy questions and students' scores in an OSPE of 2nd year medical students, conducted in a dissecting room. It also examined the relationship between students' scores and certain demography, career aspirations and participation in voluntary anatomy activities provided at the school. It extends the work of authors such as Inuwa et al. (2011, 2012) as in addition to examining the effect of different visual resources it also examines the effects of (i) the gender of the question asked in reproductive anatomy (ii) the addition of a clinical scenario based on the structure to be identified and (iii) participation in a variety of voluntary anatomy related activities i.e.: voluntary revision sessions, Summer dissection program, and Demonstrator program.

MATERIALS AND METHODS

Approvals

St. George's, University of London (SGUL) on February 25, 2013, and UCL, Institute of Education, London on March 10, 2013, ethically approved the research project. Any personal data was collected with the informed consent of the students who agreed to participate, and data collected was stored in compliance with the Data Protection Act 1998.

Permission was also sought and obtained from the undergraduate curriculum lead and the examination team to include questions with clinical scenarios in the OSPE, which was not standard practice at SGUL at the time of this study. Students were informed of the changes during teaching and were set a mock OSPE at the end of semester's teaching in order to familiarize them with the format before using it "live" at the end of the fourth semester. The examination was set up on June 27, 2013 and it was administered on June 28, 2013. No other changes were made to the established teaching and assessment practices.

Context of the study

The anatomy curriculum is in line with the core curriculum for medical students as recommended by the General Medical Council "Tomorrows Doctors" (McHanwell et al., 2007).

The students tested in the OSPE described and analyzed here were enrolled in MBBS5 a medical undergraduate stream taking 5 years in total. Years 1 and 2 contained four curriculum themes: basic and clinical sciences (BCS), patient and doctor (P&D), community and population health (CPH), and personal & professional development (PPD). These themes were tied to three assessment themes BCS, P&D and CPH/PPD. In order to progress past year 2, students were required to pass all of the above themes. Assessment was through OSPE, Objective Structural Clinical examination (OSCE) and written examinations. The anatomical component was tested by OSPE

The students were taught anatomy using Problem Based Learning (PBL) in a spiral curriculum. Semester 1 was a general overview; cardiovascular, respiratory, urinary, alimentary and endocrine systems were taught in semester 2, musculoskeletal system and neuroanatomy in semester 3 and the reproductive system (which is the focus of this study) in Semester 4. Reproductive anatomy was taught in three weeks through three lectures (pelvis and perineum, female, and male reproductive systems), three corresponding practical sessions, three formative quizzes (sent at the end of each week) and one mock OSPE at the end of the third week. One lecturer taught the three lectures, and 20 demonstrators (one at each dissecting table) supervised by an academic facilitated the dissecting room sessions. The sessions utilized a mixture of prosected specimens/ dissected cadavers, bones, plastic models and medical images, but students do not dissect (except in summer dissection program).

In the OSPE reported in this study, 75% of the questions were based on the topics covered in semester 4 (i.e. reproductive anatomy), and the rest on topics covered in semesters 1, 2 and 3. The questions analyzed in this study were concerned with reproductive anatomy and not with prior topics.

The OSPE reported here was a timed practical examination using cadaveric specimens, bones, plastic models and medical images arranged in two circuits of 30 stations run simultaneously. Each station comprised two related SBA questions. Part 1 required identification of a tagged or pinned anatomical structure on a specimen or medical image. Part 2 was intended to assess the relevant functional, applied or clinical

knowledge. A proportion of these questions were linked to a clinical case scenario which consisted of symptoms and history followed by some diagnosis information.

Each student was given 90 seconds to complete each station before being moved on to the next station in response to a buzzer signal. Touching the specimens was not permitted. The students answered the questions by putting a cross in the appropriate box on an answer sheet, which was marked electronically.

A total of 150 students were assessed in two groups one of 76 and the other of 74. As the examination was run twice with only 30 stations, the “live” stations were interspersed with “wait stations” to accommodate the extra numbers.

Experimental Design

Hypotheses

The experiment was intended to test the effect of multiple factors on the performance of students in the context of an OSPE examination. These factors were:

1. Design of Question:
    - Visual context: cadaveric resource vs. medical image
    - Clinical context: use of clinical scenario vs. no clinical scenario
    - Gender context: male reproductive anatomy vs. female reproductive anatomy
2. Student demographics
    - Age, gender and ethnicity
    - Learning preferences (cadaveric vs. online)

- Examination preferences (cadaveric vs. online)
- Anticipated career choice
- Participation in voluntary anatomy programs

Information on student demography and the student related items described above was collected by a voluntary questionnaire.

The null hypothesis in each case was that the factor had no significant effect on the student performance.

Difficulty of question was an obvious factor on question-design and so it was included as a factor in all analyses.

Design of the OSPE questions

The examination was constructed in the usual way for this institution, that is: (i) learning objectives were identified, (ii) the relevant anatomical domain and the related clinical problems for designing OSPE questions were identified. (iii) the questions were carefully sampled across the course content and (iv) the paper was reviewed by a panel of internal and external examiners in accordance with the school's quality assurance strategy – focusing on the basic rules for designing SBA questions, that is: the stem of each station posed a clear question; all distractors were homogenous and every attempt had been made to avoid technical item flaws (Case and Swanson, 2002). Distractors were arranged in alphabetical order before administration of the examination to limit cueing effects.

In accordance with standard marking practice, the 22 two-part questions relevant to this study were dissociated into 44 separate questions, which were marked independently. For the purposes of the study, the questions were ranked by a panel of three members of staff as easy, moderate or difficult using an adapted Ebel's matrix (Ebel, 1979). After the examination, the questions were re-ranked into two categories (difficult and easy) by the same panel (who were unaware of the outcome) because it was felt by the analyzing statistician that question numbers were insufficient to support a robust analysis of three question categories.

Examples of two stations are shown in Table 1 with figure 1 and figure 2.

Design of the questionnaire

The questionnaire was of a simple design, collecting student demographics and participation in voluntary anatomy related activities, there was no intent to construct scores, scales or latent variables and thus no requirement for reliability or validity testing; it was reviewed by three members of staff before use.

Voluntary anatomy activities

Students were asked to indicate their intention to participate in or actual participation in three voluntary activities. These were:

- The Summer dissection program (dissecting a half-cadaver in the summer break)
- The Anatomy Demonstrator program (senior students assisting academics in teaching their juniors)

- Voluntary revision sessions (use of dissecting room when not otherwise occupied)

The questionnaire was designed using survey-monkey software but responses were collected from a paper version, which was administered at the end of each practical session.

RESULTS

Eighty-five out of 150 students consented to complete the questionnaire study (57% of the class), 83 of these actually completed it. These students had their demography linked to their performance scores. Student Demographics from the questionnaires are shown in Table 2 (Students Demographics – Consented students only). Ethnicity was originally collected according to a standard nineteen-category scale; this was collapsed into five categories as shown, and eventually into two categories (white and other) for analysis. Age was collapsed into two categories: date of birth 1991-1993 and date of birth pre1991, this categorization was based on the fact that those born pre 1991 were old enough to have completed a prior degree or to have experience of life outside education, whereas the younger category would have moved straight from high school to university.

Marks from the students who did not consent to the questionnaire were used in the analysis of the question design part of the study, but marks were irretrievably de-identified, so it was not possible to trace individual students marks in these students.

Effect of Question Design

Analysis was carried out by multiple linear regression using SPSS v. 21(IBM SPSS Statistics for Windows, version 21.0. Armonk, NY: IBM Corp). The dependent variable was the number of students who correctly answered each question, so the maximum possible score was 150 and the minimum was zero. Regression was chosen as a means of analysis because all questions, of necessity belonged to one category of all four groups, and regression was the most suitable way of separating the effects of the four pairs of binary categories.

The total number of correct responses was used as a dependent variable. The independent variables tested were: (i) gender of question female vs. male, (ii) cadaveric resource vs. image, (iii) presence of clinical scenario vs. no scenario and (iv) easy vs. difficult. All were coded as binary categorical variables. Table 3 shows summary statistics for the question-categories.

The regression analysis showed that the only significant effect was the difficulty of the question ($p<0.001$); the adjusted mean difference in scores between difficult and easy was 69 marks (Table 3: Summary statistics for question-categories). Gender of the question was insignificant ($p=0.351$) as so was cadaveric resource vs. image ($p=0.351$) and presence/absence of clinical scenario ($p=0.716$).

Effect of demography and other questionnaire data:

Analysis was again carried out using multiple linear regression to allow adjustment for potential confounding. The dependent variable was individual student mark. Initially, the marks of the students who completed the questionnaire were compared with the marks of those who did not. Those who consented scored a mean of 40.9 marks and the group who did not consent a mean of 2.2 marks less. This difference was statistically significant ($p = 0.07$) but is considered small enough to have no contextual significance, suggesting that the students who consented are a representative sample of the whole class.

All of the other data collected in the demographic questionnaire were tested in the regression model, Initially, all variables were tested in a univariate analysis, and purposeful selection with a cutoff of $p>0.2$ was used to select variables for further modeling. The independent variables selected for further modeling were: Interest in anatomy demonstrator program ($p=0.001$), attended voluntary revision sessions ($p=0.182$), ethnicity collapsed to white vs. other ($p=0.184$) and student gender ($p=0.104$). Summer dissection program ($p=0.825$), age ($p=0.812$) and career choice ($p=0.787$) were discarded. The final best-fit model showed only participation in demonstrator program as a significant predictor of mark ($p<0.001$) with a trend toward significance for student gender ($p=0.051$), the adjusted mean difference between male and female students was 4.4 marks, with the mean females' mark being higher.

Effect of Two-part questions

A binary logistic regression was performed with Part 2 score as the dependent variable and Part 1 score as the independent variable to determine whether a correct answer to Part 1 could predict whether a student would answer Part 2 correctly. When tested alone, a correct answer to Part 1 was highly predictive of a correct part 2 answer ($p<0.001$). However, there was an uneven distribution of easy and difficult questions between part 1 and part 2; part 1 contained 16 “easy” questions and 6 “difficult” ones, the ratio was reversed in part 2. When adjusted for difficulty, a correct answer for Part 1 was no longer predictive of a correct answer to part 2 ($p=0.765$).

DISCUSSION

This study is the first of its kind to investigate the effect of the design of OSPE questions on the performance of 2nd year medical students. The principle finding was that only question difficulty was a significant predictor of the performance of a group sample of students; gender of question, presence/absence of a clinical scenario and the type of visual resource presented all had no effect.

In this study the only factor that was a significant predictor of performance was interest or participation in the demonstrator program, which is to be expected, since it is likely that those students expressing interest will be the most motivated and interested in anatomical subjects. In anatomy, teaching junior students/demonstrating methods are often used in a number of medical schools, and are known to provide an effective learning environment (Evans and Cuffe, 2009).

Participation in the voluntary revision sessions or in the summer dissection programs had no effect on the mean mark. However, the demographics (Table 2) show that the majority of students (74/83) attended voluntary revision sessions, so this finding has limited reliability as asymmetric distributions in a binary variable result in very wide confidence intervals in a regression analysis. The same is true of learning preferences and assessment preferences, 89% and 92% respectively favored dissection room approaches, so these results must also be considered of limited reliability. It would be expected however that the voluntary summer dissection class would attract those with more ability and interest in anatomy, although it is equally possible that it may attract those weak in anatomy who hope to extend their knowledge by participating; and this may be why the summer dissection program is insignificant, it is attracting the best and the worst. Anticipated Career choice (surgical, non-surgical, don't know) also showed no effect on performance, but since 44.7% did not express a preference, the power to detect differences is low.

So far as is known, there is no prior work or any theoretical framework suggesting that ethnicity is a factor in anatomy learning and there was no effect seen here. The effect of age was tested because it was hypothesized that older students with prior degrees or equivalent life experience might be more effective learners and score more highly, however, only 6 out of 83 students were of an age to have done anything other than move from high school to medical school, so this must also be treated as an unreliable finding.

Students' gender as an effect on anatomical learning has been studied before, and Hisley et al. (2007), made similar finding to this study that is a trend towards significance with females performing better than males in a study comparing dissection with digital dissection.

The students' better scores on cadaveric questions could be explained by their better understanding of actual three-dimensional resources as compared to two-dimensional radiological images that represent three-dimensional structures. It could also be explained by relatively less use of two-dimensional resources in previous assessment tools and teaching sessions or possibly their lack of ability of appropriately linking knowledge to read two-dimensional grey scale medical images (Miller, 2000). However, several authors have found no significant differences in performance of groups assessed with factual questions that test the recall of knowledge, compared to groups assessed on various visual resources that require understanding and interpretation (Khalil et al., 2005; Schubert et al. 2009; Inuwa et al., 2011, 2012), so possibly this finding cannot be attributed to inexperience in interpreting 2D images.

Incorporating clinical scenario/case histories in anatomy teaching and assessments has often been advocated in the literature (Moxham et al. 2011; Yaqinuddin et al. 2013); but information on students' performance on questions with and without scenarios is scarce. The lack of effect on mark in this study is reassuring and suggests that clinical scenario has a place in assessment as well as teaching. In principle it could uplift the level of

knowledge assessed from “knows” to “knows how” (Miller, 1990) and measure levels 1, 2 and 3 simultaneously of modified Bloom’s taxonomy (Bloom, 1956; Palmar and Devitt, 2007). It is possible that their use could steer students’ future learning patterns towards applied and clinical anatomy.

Of the four factors examined in question design, (gender, scenario vs. no scenario, cadaveric resource vs. image, easy vs. difficult), only difficulty was a significant predictor of the number of correct answers. This is a reassuring finding, suggesting that at least in this instance the OSPE was a valid test of student knowledge and the results were not biased by other factors.

In an OSPE with two part questions such as the one reported here, the goal is to compose questions such that failure to answer the first part correctly does not handicap the student in answering the second part. It is debatable how successfully this can be achieved in two parts questions and until now there have been no analyses as to whether this has been achieved in practice. In this case, a correct answer to part 1 was not predictive of a correct answer to part 2, any differences in Part 1 vs. Part 2 marks were attributable to imbalances in the difficulty.

LIMITATIONS

This study has several limitations, the sample size was fixed by the size of the class in that year, the demography and learning assessment preferences could only be observed in the consented group, and in reality there were several factors where there

was little chance of achieving reliable comparisons between two categories because of asymmetric distribution between them. Additionally, the study is limited to one facet of anatomy and may not be generalizable to other areas without more study and accumulation of larger samples. As it stands, this is a study from one year and one medical school, which indicates some directions for further study.

CONCLUSION

None of the question designs tested in this study were significantly related to the number of correct answers; only difficult was a significant predictor. Of the various optional activities offered to students in this medical program, only interest or participation in the demonstrator program was a significant predictor of high scores in the OSPE. It appears that in this OSPE, Part 2 questions could be correctly answered without necessarily answering Part 1 correctly.

ACKNOWLEDGEMENTS

The authors acknowledge the work of colleagues in Biomedical Sciences (Anatomy) and the examination team at the school, and Dr. Mothana Gawad, Dr. Richard Wingate, Dr. Charlie Owen, Dr. Caroline Pelletier and Professor Harold Ellis.

NOTES ON CONTRIBUTORS

MANDEEP GILL SAGOO, M.Sc. (Med Anat), P.G.C.E., is a fellow of the Higher Education Academy, anatomy teaching fellow in the Department of Anatomy and

Human Sciences at King's College, London, and final-year doctorate in education student at UCL, Institute of Education, London, United Kingdom.

CLAIRE FRANCE SMITH, B.Sc., P.G.C.E., Ph.D., F.H.E.A., F.A.S., is reader in Anatomy and Head of Anatomy at Brighton and Sussex Medical School, University of Sussex, Brighton East Sussex, United Kingdom. Her research is in understanding the learning experience including: approaches to learning, learning psychometrics and spatial ability. She is also Chair of Education Committee for the Anatomical Society of Great Britain.

EDWARD GOSDEN, M.Sc., is a statistical epidemiologist in the Institute of Health and Biomedical Innovation at Queensland University of Technology, Brisbane, Australia. His research interest is in epidemiology and the analysis of clinical trials with regression modelling.

LITERATURE CITED

Bloom BS.1956. Taxonomy of Educational Objectives: the classification of educational goals. New York: David McKay Co. Inc.

Case SM, Swanson DB. 2002. Constructing Written Test Questions for the Basic and Clinical Sciences. 3rd Edition. Philadelphia, PA: National Board of Medical Examiners. 180 p. URL: http://www.nbme.org/pdf/itemwriting_2003/2003iwgwhole.pdf [accessed 20 March 2015].

Clough RW, Lehr RP. 1996. Testing knowledge of human gross anatomy in medical school: an applied contextual-learning theory method. Clin Anat 9(4):263-268.

Daly FJ. 2010. Use of electronic anatomy practical examinations for remediating “at risk” students. Anat Sci Edu 3:46-49.

Ebel RL 1979. Essentials of Education Measurement. 3rd Edition. Upper Saddle River, NJ: Prentice Hall Inc. 388p.

Evans DJ, Cuffe T. 2009. Near-peer teaching an anatomy: an approach for deeper learning. Anat Sci Edu 2(5):227-233.

Fabrizio PA. 2013. Oral anatomy laboratory examinations in a physical therapy program. Anat Sci Edu 6:271-276.

Fitzgerald JE, White MJ, Tang SW, Maxwell-Armstrong CA, James DK. 2009. Are we teaching sufficient anatomy at medical school? The opinions of newly qualified doctors. Clin Anat 22:288–289.

Gunderman RB. 2008. Giving ourselves: The ethics of anatomical donation. Anat Sci Educ 1:217–219.

Hisley KC, Anderson LD, Smith SE, Kavic SM, Tracey JK. 2008. Coupled physical and digital cadaver dissection followed by a visual text protocol provides insights into the nature of anatomical knowledge and its evaluation. Amer Assoc Anat 1:27-40.

Inuwa IM, Taranikanti V, Al-Rawahy M, Habbal O. 2011. Perceptions and attitudes of students towards two methods of assessing practical anatomy knowledge. Sultan Qaboos Univ Med J 11:383–390.

Inuwa IM, Taranikanti V, Al-Rawahy M, Habbal O. 2012. Anatomy practical examinations: How does student performance on computerized evaluation compare with the traditional format? Anat Sci Educ 5:27–32.

Khalil MK, Johnson TE, Lamar CH. 2005. Comparison of computer-based and paper-based imagery strategies in learning anatomy. Clin Anat 18:457–464.

Krippendorf BB, Bolender DL, Kolesari GL. 2008. Computerized grading of anatomy laboratory practical examinations. Anat Sci Educ 1:220–223.

McHanwell S, Atkinson M, Davies DC, Dyball R, Morris J, Ockleford C, Parkin I, Standring S, Whiten S, Wilton J. 2007. A core syllabus in anatomy—adding common sense to need to know. Eur J Anat 11:S3–S18.

Menezes RG, Nayak VC, Binu VS, Kanchan T, Rao PP, Jagdish RP, Baral P, Lobo SW. 2011. Objective structural practical examination in forensic medicine: students' point of view. J Forensic Leg Med 18:347-349.

Miller GE. 1990. The assessment of clinical skills/competence/performance. Acad Med 65:63–67.

Miller R. 2000. Approaches to learning spatial relationships in gross anatomy: Perspective from wider principles of learning. Clin Anat 13:439–443.

Moxham BJ, Shaw H, Crowson R, Plaisant O. 2011. The future of clinical anatomy. Eur J Anat 15:29–46.

Nayar U, Malik SL, Bijlani RL. 1986. Objective structured practical examination: A new concept in assessment of laboratory exercises in preclinical sciences. Med Educ 20:204–209.

Palmer EJ, Devitt PG. 2007. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple-choice questions? BMC Med Edu 7:49.

Regan de Bere S, Mattick K. 2010. From anatomical 'competence' to complex capability. The views and experiences of UK tutors on how we should teach anatomy to medical students. Adv Health Sci Educ Theory Pract 15:573–585.

Rowland S, Ahmed K, Davies DC, Ashrafian H, Patel V, Darzi A, Paraskeva PA, Athanasiou T. 2011. Assessment of anatomical knowledge for clinical practice: Perceptions of clinicians and students. Surg Rad Anat 33:263–269.

Schnotz W, Baadte C. 2015. Surface and deep structures in graphics comprehension. Mem Cogn 43:605-618.

Schubert S, Schnabel KP, Winkelmann A. 2009. Assessment of spatial anatomical knowledge with a 'three-dimensional multiple choice test' (3D-MC). Med Teach 3:13–17.

Smith CF, Mathias HS. 2011. What impact does anatomy education have on clinical practice? Clin Anat 24:113–119.

Smith CF, McManus B. 2015. The integrated anatomy practical paper: A robust assessment method for anatomy education today. Anat Sci Educ 8:63–73.

Vorstenbosch MA, Kooloos JG, Bolhuis SM, Laan RF. 2015. An investigation of anatomical competence in junior medical doctors. Anat Sci Educ (in press; doi: 10.1002/ase.1513).

Yaqinuddin A, Zafar M, Ikram MF, Ganguly P. 2013. What is an objective structured practical examination in anatomy? Anat Sci Educ 6:125–133.

TABLES

Tables 1-3 (shown on another document)