Online Assessment of Applied Anatomy Knowledge: The Effect of Images on Medical Students' Performance

Mandeep Gill Sagoo, [iD] [1]* Marc A.T.M. Vorstenbosch,[2] Peter J. Bazira,[3] Harold Ellis,[1] Maria Kambouri,[4] Charlie Owen[5]

[1]Department of Anatomy, School of Centre for Education, King's College London, London, United Kingdom
[2]Department of Anatomy, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands
[3]Centre for Anatomical and Human Sciences, Hull York Medical School, York, United Kingdom
[4]Department of Psychology and Human Development, Institute of Education, University College London, London, United Kingdom
[5]Thomas Coram Research Unit, Institute of Education, University College London, London, United Kingdom

Anatomical examinations have been designed to assess topographical and/or applied knowledge of anatomy with or without the inclusion of visual resources such as cadaveric specimens or images, radiological images, and/or clinical photographs. Multimedia learning theories have advanced the understanding of how words and images are processed during learning. However, the evidence of the impact of including anatomical and radiological images within written assessments is sparse. This study investigates the impact of including images within clinically oriented single-best-answer questions on students' scores in a tailored online tool. Second-year medical students ($n = 174$) from six schools in the United Kingdom participated voluntarily in the examination, and 55 students provided free-text comments which were thematically analyzed. All questions were categorized as to whether their stimulus format was purely textual or included an associated image. The type (anatomical and radiological image) and deep structure of images (question referring to a bone or soft tissue on the image) were taken into consideration. Students scored significantly better on questions with images compared to questions without images ($P < 0.001$), and on questions referring to bones than to soft tissue ($P < 0.001$), but no difference was found in their performance on anatomical and radiological image questions. The coding highlighted areas of "test applicability" and "challenges faced by the students." In conclusion, images are critical in medical practice for investigating a patient's anatomy, and this study sets out a way to understand the effects of images on students' performance and their views in commonly employed written assessments. Anat Sci Educ 14: 342–351. © 2020 The Authors. Anatomical Sciences Education published by Wiley Periodicals LLC on behalf of American Association of Anatomists.

Key words: gross anatomy education; medical education; undergraduate education; single-best answers; applied anatomy; visual psychology; anatomy imaging; online assessment; schemas

INTRODUCTION

Anatomical knowledge is essential for physical examination, interpreting radiological images, establishing a working diagnosis, carrying out clinical procedures, performing surgical procedures, and understanding anatomical pathology (Older, 2004; McHanwell et al., 2007; Dettmer et al., 2013; Orsbon et al., 2014; Vorstenbosch et al., 2016). With increasing expertise, this knowledge becomes encapsulated in clinical concepts and used more implicitly in clinical reasoning (Boshuizen and Schmidt, 1992; Schmidt and Rikers, 2007).

Despite its clinical relevance, anatomy teaching and assessment have been the subject of considerable debate because of pressure from competing space and time demands within

curricula and institutions (Moxham et al., 2011). However, there is a significant tendency toward educational approaches that facilitate the application of knowledge in practice (McHanwell et al., 2007; Ahmed et al., 2010).

The majority of the literature on assessment methods in anatomy typically addresses assessment utility indices such as validity, reliability, and educational impact (van der Vleuten and Schuwirth, 2005; Samarasekera et al., 2015), or the pedagogic influence of visual resources in factual multiple-choice questions (Khalil et al., 2005; Inuwa et al., 2011, 2012). Anatomy assessments typically test factual and/or applied anatomy knowledge with or without the inclusion of visual resources. Furthermore, there has been an increasing demand for junior doctors to have detailed knowledge of imaging anatomy, and this emphasizes the multifaceted nature of the subject beyond cadaveric anatomy, that is, understanding different types of images and cross-sections (Phillips et al., 2013). However, it leaves the impact of images in clinically oriented questions unarticulated (Phillips et al., 2013). It is widely acknowledged that such images provide a powerful learning stimulus and help medical students understand anatomy both in health and disease (McHanwell et al., 2007), and the ability to acquire adequate visual internal representations of anatomical information is an essential element of learning anatomy (Vorstenbosch et al., 2016). For the majority of doctors, such images are the main representations of internal anatomy utilized in clinical practice and provide an intrinsic, built-in meaning distinct from that of diagrams and illustrations in anatomical texts (Schnotz, 2002). This raises the critical question of how to educate our medical students adequately, and equip them to understand medical images.

The majority of research on the use of images in learning is based on recognition memory, the transfer of learning content from images to text and vice versa (Ginns, 2005; Witteman and Segers, 2010) or on the use of images as a motivational benefit for learners (Ainsworth, 1999). However, the Cognitive Theory of Multimedia Learning (CTML) suggests that people learn better from a combination of words (spoken or written) and images (illustrations, photos, animation, or videos) than from either words or images alone (Biedermann, 1981; Mayer, 2005a; Mayer, 2009). Cognitive Theory of Multimedia Learning is based on three key assumptions: the "dual-channel assumption," the "limited capacity assumption" and the "active processing assumption." These assumptions and the instructional principles based on them draw on Paivio's (1986) dual coding theory, Baddeley's (1992) model of working memory, and Sweller's (1994) cognitive load theory.

Paivio's dual coding theory depicts the human cognitive system as dependent on verbal and imagery subsystems. In this respect, integrative processing through referential connections is thought most likely to occur if verbal and visual information are simultaneously available in working memory. Similarly, Baddeley's model of working memory (Baddeley, 1992) posits the existence of auditory and visual subsystems in limited working memory. The dual-channel assumption of CTML merges Baddeley's and Paivio's conceptions, positing that humans process information in working memory through two channels: an auditory–verbal channel and a visual–pictorial channel. The second assumption of CTML, reflecting both the work of Baddeley (1992) and Chandler and Sweller (1991), is that these two channels have a limited capacity to convey and process information. The third assumption of CTML is that humans are active sense-makers; engaging in active cognitive processing to construct coherent knowledge structures, or

"schemas," from a combination of prior knowledge and external information. "Schemas" are meaningful sets of connections that correspond to specific concepts and experiences, and the acquisition of expertise in any area can be characterized by the development of this idiosyncratic memory (Regehr and Norman, 1996). In medical education, for instance, schemas are used to combine a variety of isolated facts, aggregate these into concise and dense "illness scripts," which are then enriched by experience into "instance scripts" (Schuwirth and van der Vleuten, 2011). These instance scripts enable the instantaneous recognition of disease patterns by experts.

Although it is widely accepted that multiple representations of information can complement or support learning (Ainsworth, 1999), the parallels drawn between text processing and image processing have been questioned by Schnotz and Bannert (2003). Advancing their theory of Alternative Multimedia learning, they argued that the use of visual resources such as images in learning comes with both cognitive benefits and cognitive costs. However, there are significant similarities in the processing of text and images. Like textual information, images possess both perceptual surface structure and a deep semantic structure (Schnotz and Baadte, 2015). The surface structure of an image includes dots, lines, areas, and their visual features whereas the deep structure of an image is a semantic construct which expresses its meaning. Making sense of, or "processing" an image or text requires the leveraging of prior knowledge to integrate the external representation (the image or text) with its internal semantic representation, and thus requires the use of schemas for comprehension. What this literature suggests is that understanding image is a matter of complex interactions between several factors including perceptual surface structures, deep semantic structures, and association and inference with cognitive schema (Crisp and Sweiry, 2006; Schnotz and Baadte, 2015).

The idea therefore that images, in general, improve learning has been widely accepted; however, in assessment, the role and impact of images are ambiguous (Schuwirth and van der Vleuten, 2004a, b; Crisp and Sweiry, 2006; Schnotz and Baadte, 2015). Generalizing the role of images in assessments in subjects like anatomy adds another level of complexity because the process of learning anatomy is considered as learning from images supported by text, rather than vice versa. Additionally, these images require pre-existing knowledge to interpret (Schnotz, 2002). The cognitive domain of assessment is categorized into "knowledge/ content dimension" and "cognitive process/progress dimension." In anatomy, the content dimension includes anatomical terminology (with associated images) and facts, conceptual and procedural knowledge. In contrast, the progress dimension demonstrates the understanding of facts, ideas, and images by organizing, comparing, interpreting, and applying the knowledge gained (Brenner et al., 2015). Owing to the need for authenticity and face validity, i.e. the extent to which a test is compatible with its educational philosophy (van der Vleuten and Schuwirth, 2005; Gunderman, 2008; Sugand et al., 2010; Samarasekera et al., 2015), students are under pressure to develop schemas for relevant text and visuals, and simultaneously be capable of interpreting these visuals used in anatomy and clinical settings.

Several studies have investigated responses to various types of images in medical assessments. These include studies on extended matching questions (EMQs) with labeled images versus textual material (Vorstenbosch et al., 2013, 2014); multiple-choice questions (MCQs) with images versus textual description of images (Hunt, 1978); spotter test with cadaveric specimens versus online resources (Inuwa et al., 2011, 2012); identification questions with online interactive

images, static line diagrams versus real objects (Khalil et al., 2005); MCQs with cadaveric and textual material (Schubert et al., 2009); MCQs with simplistic diagrams versus histology images (Holland et al., 2015); MCQs with and without images (Notebaert, 2017); postgraduate surgical MCQs with images versus verbal questions (Buzzard and Bandarnayake, 1991), and illustrated and nonillustrated MCQs in a medical licensing examination (Bahlmann, 2018). Some of these studies showed consistent positive, negative, or no effects (Berends and Van Lieshout, 2009; Holland et al., 2015) whereas others showed inconsistency (Hunt, 1978; Vorstenbosch et al., 2013) in students' performance and preferences. These studies are mainly focused on factual identification type questions (Khalil et al., 2005; Inuwa et al., 2011, 2012).

Although students and teachers appear to prefer visual resources in anatomy (Older, 2004; Rowland et al., 2011; Orsbon et al., 2014), the effects on the performance of clinically oriented anatomy questions with and without images are thus far inconclusive. This is reflected in the fact that best-practice guidelines on question writing (Case and Swanson, 2002; Wood et al., 2004) give no explicit guidance about the use of images in writing multiple-choice questions to test application of basic science knowledge.

This study, therefore, aimed to investigate how the inclusion of images in clinically oriented anatomy assessment affects student performance. The objective of this study was to answer the following research question: What is the effect of purely textual and image-based clinically oriented single-best-answer questions on student performance in an anatomy examination and on their views derived from their free-text comments? The study hypothesized that the use of anatomical/radiological images in questions would have a positive effect on the students' performance compared to text-only questions.

MATERIALS AND METHODS

Medical Schools' Involvement

"Standard entry medical students" from six medical schools in the United Kingdom participated in the study (MSC, 2020). These schools use a variety of available anatomical resources in their curricula, including but not limited to cadaveric dissections, prosections, and/or radiological images as shown in Table 1.

Participants

Participants ($n = 174$) were medical students at the end of their preclinical year, i.e., end of the second year, having completed the formal anatomy curriculum and due to take their final summative assessments in anatomy. The test was reviewed by the academic leads of anatomy from the participating medical

Table 1.

Anatomy Teaching Resources, Delivery Formats, and Types of Assessments in Participating Schools

School	Teaching resource	Delivery format	Approximate number of hours	Assessment type
1	• Dissection • Prosected specimens • Bones • Plastic models • Images	• Lectures • Practical classes • Living anatomy	90	• MCQ • SAQ • Formative *viva*
2	• Prosected specimens • Bones • Plastic Models • Images	• Lectures • Practical classes • Living anatomy	80	• Spotter • SBA • MCQ • OSCE
3	• Dissection • Prosected specimens • Bones • Plastic models • Images	• Lectures, • Practical classes • Living Anatomy	140	• SBA • OSCE
4	• Images, • Plastic models • Virtual dissection	• Problem-based anatomy sessions • Clinical skills	100	• Progress test • Anatomy spotter
5	• Prosected specimens • Bones • Plastic models • Images	• Lectures practical Classes • Living anatomy	110	• OSPE • SBA
6	• Virtual dissection • Images • Plastic models	• Integrated Anatomy/Pathology sessions • Living anatomy	90	• Progress test

MCQ, Multiple-choice questions; SAQ. Short-answer questions; Formative *viva,* oral examination for learning; SBA, Single-best answers; OSCE, Objective structured clinical examinations; OSPE, Objective structured practical examination. Progress tests utilize single-best-answer questions to assess students' knowledge against graduate learning outcomes.

Sagoo et al.

schools to confirm the equivalence of knowledge and homogeneity of the students.

Bespoke Online Assessment Tool

A bespoke online assessment tool "My Anatomy Growth" was built and coded by a professional software programmer and securely hosted on Microsoft Azure Cloud Services (Microsoft Corp, Redmond WA; Windows Server 2012 R2, North Europe). Eight clinically qualified anatomy academics reviewed the content, and the assessment tool piloted in seven second-year medical students.

The decision to utilize "My Anatomy Growth" followed a rigorous evaluation of existing online assessment and survey tools, such as Storyline 1, version 8.0 (Articulate Global, Inc., New York, NY), Perception, version 5.2 (Questionmark Computing Ltd., London UK), GoogleForms, version 0.8 (Google LLC, Mountain View, CA), Opinio, version 6.8 (University College London, London UK), and SurveyMonkey, version 2014 (SurveyMonkey®, San Mateo, CA). None of these existing tools met all the requirements for the study. The requirements were cross-browser compatibility, a precise and customized look; an efficiently secured system, data registration and log-in through authenticated email addresses; the ability to incorporate the participant information sheet, consent form, and collect demographic data before presenting the test; the ability to randomize questions within the test and to start the clock and allocate 1 hour 30 minutes to complete the test (see Supporting Information).

Comprehensive formative feedback on each answer and distractor for all questions was provided to students upon completion of the test. Furthermore, a free-text comment box was added on the last page of the test, along with a "thank you" note for their participation.

Ethical Considerations

Ethical clearance was sought and received from each of the six participating medical schools as follows: 30 March 2015 (School 5), 9 April 2015 (School 1), 14 April 2015 (School 3), 17 April 2015 (School 2), 20 April 2015 (School 4), and 25 June 2015 (School 6).

Participants registered and logged in with their unique institutional authenticated email address, and were presented with the "participant information sheet" before documenting their informed consent. Both the "participant information sheet" and "consent form" were designed to the British Educational Research Association (BERA) guidelines (BERA, 2018).

Study Design

A quasi-experimental design was employed with the medical schools being selected, and the participants (medical students) independently volunteering to take part in the study. The same online test was taken, and students' scores on the question types were analyzed. Upon completion of the test, although unprompted, 55 students provided free-text comments on the tool and the design of the questions, and these were thematically analyzed.

Design of the test questions. A total of thirty-six questions were thematically organized with an equal distribution of 12 questions each covering the following three anatomical regions:

limbs (lower and upper limbs), torso (thorax, abdomen, and pelvis), head and neck (including brain and vertebral column). These twelve regional anatomy questions comprised four nonimage questions, four anatomical image questions, and four radiological image questions.

For validation, the test was blueprinted to the Anatomical Society and the General Medical Council's "*Tomorrow's Doctors*" criteria (McHanwell et al., 2007; GMC, 2009; Louw et al., 2009; Biggs and Tang, 2011).

Each question in the test was explicitly linked to a specific anatomical domain and clinical relevance and designed according to best-practice guidelines for the development of single-best items (Haladyna and Rodriguez, 2013). The textual information in each question was unique (Fig. 1 shows examples of questions with and without images). The schematic diagram of the question distribution is shown in Figure 2.

Furthermore, students' performance on questions with images indicating soft tissue and bones was carried out because anatomical and radiological images are not homogenous images, i.e., bones appear different from soft tissue in these images. In this study, the surface structure refers to the type of image (anatomical and radiological image), and deep structure is a semantic construct which expresses the meaning of the image, i.e., what it is that students are required to conceptualize to answer a question worded around a bone or soft tissue in an image.

This study followed classical test theory, treating the observed score as a combination of the true score and an error score. The true score is the hypothetical score a student would obtain based on their competence. However, as every test induces measurement errors, the observed score may not necessarily be the same as the true score (Engelhardt, 2009). The reliability of this test was investigated through Cronbach's alpha (Cronbach, 1951).

Based on test scores, the participants were categorized as high performing students (achieving raw test scores of 23-34) and low performing (achieving raw test scores of 11-22) students. The test was released before their final examinations, and it was anticipated that mainly keen or borderline students would be interested in this revision tool. Hence these categories were made based on their scores on the test.

Statistical analysis of scores. The normality of the data was investigated. This was followed by using the analysis of variance (ANOVA) using SPSS statistical package for Mac computers, version 22.0 (IBM Corp., Armonk, NY). The repeated measure ANOVA was chosen to measure the difference in mean test scores (dependent variable) with different independent variables (i.e., three question types along with gender and mean scores of the students of six schools) with a level of statistical significance of $P < 0.05$ (Robson, 2011).

Changes in mean test scores and standard deviations (SD) in the three question types (no image, radiological image, and anatomical image) were analyzed. The score on each question was a discrete (categorical) variable, i.e., 0 for an incorrect answer and 1 for the correct answer; however, the total scores, means, and standard deviations were a continuous dependent variable. Pearson's correlation coefficient was used to show a measure of the strength of a linear association between any two of the variables studied.

Mauchly's test of sphericity assumption was met in this parametric test when the variances of the differences between all combinations of question types were roughly equal. The

A man is hit by a baseball on the side of his head. He immediately loses consciousness, wakes up momentarily and then passes out. He is rushed to the emergency department and is immediately scanned. The scan shows a skull fracture at the site indicated in the image. He is rushed to surgery where he undergoes burr hole decompression to relieve the pressure. After a few hours, he regains consciousness. Which of the following best describes the hemorrhage from the fracture?
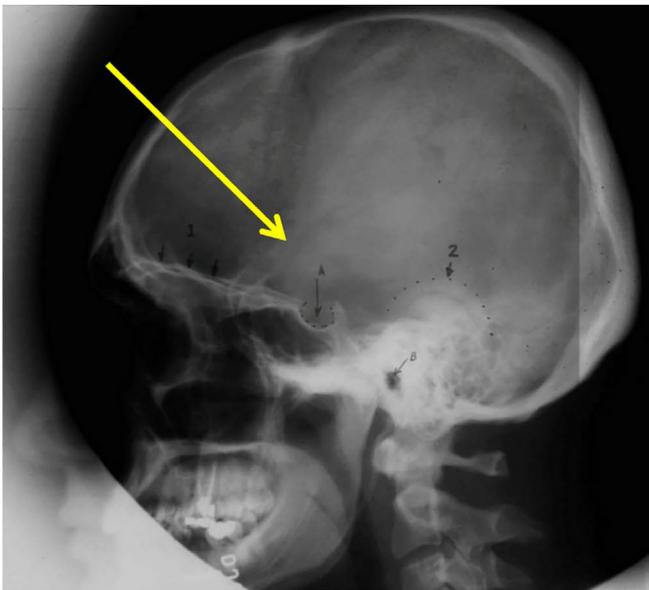
A. Extradural
B. Intracerebral
C. Subaponeurotic
D. Subarachnoid
E. Subdural

A

A 69 year old male is diagnosed with metastatic cancer. Thereafter, a secondary malignant brain tumour is found. However, despite best efforts the patient subsequently dies. An autopsy reveals tumour sites in the area indicated in the image, the vertebral column and brain but no other organs. Which of the following structures caused the cancerous cells to reach the brain?
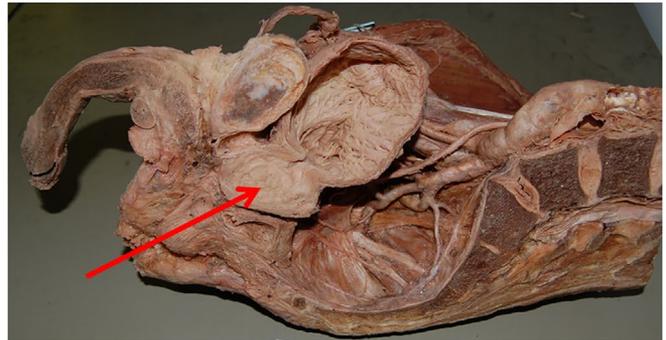
A. Anterior spinal artery
B. Azygos venous system
C. Thoracic duct
D. Vertebral artery
E. Vertebral venous plexus

B

A 52 year old woman underwent a right-sided radical mastectomy and excision of all axillary lymph nodes on the affected side due to breast cancer. Postoperatively, the nurse noticed that she had 'winging of the scapula'. Which of the following nerve damage has led to the condition?
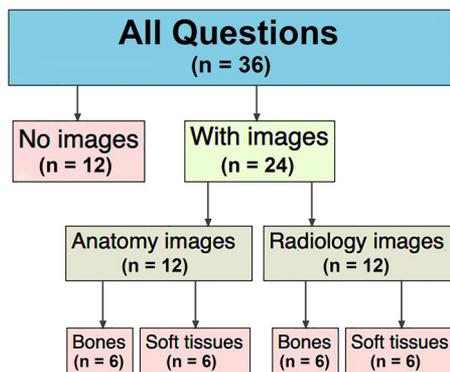
A. Axillary nerve
B. Medial pectoral nerve
C. Musculocutaneous nerve
D. Long thoracic nerve
E. Lower subscapular nerve

C

Figure 1.

Examples of clinical anatomy questions containing images. A, Question with radiology image (based on Wilson et al., 2012). Correct answer (A); B: Question with anatomy prosection image. Correct answer (E); C, Clinical anatomy question without an image. Correct answer (D).

tests of within subjects indicated the significance of the difference but did not clarify the direction of the effect. Pairwise comparisons, conducting multiple paired t-tests of scores with a Bonferroni correction to keep Type 1 error at 5% overall, were carried out to clarify the direction and size of the effect. The estimated marginal means were calculated by the ANOVA regression equation, which is the mean response from each factor, adjusted for any other variables in the model.

Measures of effect size in ANOVA are measures of the degree of association between the effect and the dependent variable. If the value of the measure of association is squared, it can be interpreted as the proportion of variance in the dependent variable that is attributable to each effect. The eta squared is a measure of effect size in ANOVA, and the partial eta squared used in this study is a proportion of variance accounted for by an effect, i.e., partial eta squared = SS effect/ (SS effect + SS error) where SS is "sums of squares," the amount of dispersion in scores. In the literature, $0.01 \leq$ partial eta squared < 0.06 is considered as small effect, $0.06 =$ partial eta squared < 0.14 is considered as medium effect, and partial eta squared ≥ 0.14 is large effect (Robson, 2011).

Analysis of free-text comments. These free-text comments on the tool and the design of the questions provided by 55 students were carefully read, processed, and organized into codes by two independent reviewers (King and Horrocks, 2010). The codes were then revised, and their inter-rater reliability checked using SPSS statistical package (Landis and Koch, 1977).

Figure 2.

Schematic diagram of the questions showing distribution across question types. A total of 36 questions were thematically organized with an equal distribution of 12 questions each covering the following three anatomical regions: limbs (lower and upper limbs), torso (thorax, abdomen, and pelvis), head and neck (including brain and vertebral column).

Table 2.

Scores of Students Across Six Participating Schools with High/Low Performing Group Distribution

School	Number of students (n)	Number of participated students n (%)	Number of low performing students (n)	Number of high performing students (n)	Examination score mean % (± SD)
1	130	3 (2.3)	3	0	51.0 (± 2.52)
2	140	13 (9.3)	11	2	53.4 (± 3.06)
3	350	121 (34.6)	41	80	68.0 (± 4.85)
4	90	12 (13.3)	8	4	59.2 (± 2.57)
5	250	17 (6.8)	13	4	58.0 (± 4.78)
6	130	8 (6.2)	2	6	63.0 (± 3.88)
Total	1090	174 (16.0)	78	96	65.0 (± 4.87)

RESULTS

Quantitative Data (Students' Scores)

Cronbach's alpha of the test had an acceptable measure of 0.73. Out of 174 students, 96 were female, and 78 were male. The mean scores of females were 22.84 ± 4.67 and those of males were 23.96 ± 5.07. The age was categorized into two groups, and there were 155 students in the age group 16-24, and 19 students above 25 years. The mean scores of age group 16-24 were 23.17 ± 4.86 and those of above 25 were 24.79 ± 4.79. The mean performance across the six schools was variable, as shown in Table 2. Out of 174 students, 78 were low performing and 96 were high performing, and there were more high performing students from School 3 as compared to the other five schools. A Shapiro-Wilk's test $(P > 0.05)$ and a visual inspection of histograms, normal Q–Q plots and box plots showed that the test scores were approximately normally distributed for students' sex, age, and the schools (Shapiro and Wilk, 1965; Doane and Seward, 2011).

This was followed by investigating the effect of three question types on students' scores, as shown in Table 3. This showed a significant correlation in the performance of students in questions with and without images. Students that performed better in text-only questions also performed better on anatomical image questions (Pearson's correlation = 0.50; $P < 0.001$) and radiological image questions (Pearson's correlation = 0.45; $P < 0.001$). Mauchly's test of sphericity assumptions was met. Tests of within-subject effects and contrasts showed a significant difference in students' performance on three question types, $F (2, 344) = 12.24$, $P < 0.001$, partial eta squared = 0.07, indicating a medium effect size. Pairwise comparisons showed the scores on anatomical image questions were significantly better than no image questions; and better on radiology image questions than no image questions. However, there was no significant difference in scores on anatomical and radiology image questions.

Further analysis was carried out between image question subtypes (on bones and soft tissues) as shown in Table 4 to investigate the representation principle of the Alternative Multimedia learning (Schnotz and Bannert, 2003; Schnotz and Baadte, 2015) for the effect of the deep structure of an image on students' scores. Mauchly's test of sphericity assumptions was met. Tests of within-subject effects and contrasts showed a significant difference in the above question subtypes, $F (1, 172) = 277.31$, $P < 0.001$, partial eta squared = 0.62 indicating a very large effect size. Pairwise comparisons showed that students performed significantly better on questions referring to bones than to soft tissues regardless of the image type.

Qualitative Data (Students' Free-Text Comments)

The codes extracted from the students' 55 free-text comments were "challenging in general," "useful," "good clinical and practical context," "to incorporate in the curriculum," "anatomical prosection images difficulty," "radiology images difficulty," "preference for items with images," "clinical context too complex," and "technical feedback on the tool." These codes were arbitrarily assigned the numbers 1–9 and level of agreement between two raters judgments calculated. Cohen's Kappa

Table 3.

Students' Scores on Three Question Types

Question type	Number of questions (n)	Examination score: mean % (± SD)	Comparison of means			
			Comparator	Difference	Standard error	P-value
No image	12		Anatomy image	5.80	0.819	<0.001
Anatomy image	12	67.1 (± 1.79)	Radiology image	−0.80	0.788	>0.050
Radiology image	12	66.3 (± 2.06)	No image	−5.00	0.870	<0.001

Table 4.

Students' Scores on Image Questions and Question Subtypes

Question type	Number of question (n)	Examination score: mean % (± SD)	Comparison of means			
			Comparator	Difference	Standard error	*P*-value
Anatomy image bone	6	79.5 (± 1.00)	Anatomy image soft tissue	−24.90	0.679	<0.001
			Radiology image bone	−2.90	0.663	>0.050
Anatomy image soft tissue	6	54.6 (± 1.33)	Radiology image bone	22.00	0.754	<0.001
			Radiology image soft tissue	1.20	0.751	>0.050
Radiology image bone	6	76.6 (± 1.28)	Radiology image soft tissue	−20.80	0.736	<0.001
Radiology image soft tissue	6	55.8 (± 1.27)	Anatomy image bone	23.70	0.660	<0.001

showed a high level of agreement between the two raters' judgments, κ = 0.82, *P* < 0.0001 (Landis and Koch, 1977).

These codes were then grouped under the overarching themes: "test applicability/quality" and "challenges." "Test applicability/quality" covered the codes "useful," "good clinical and practical context," "preference for items with images," and "technical feedback on the tool." These emphasized the usefulness of contextual questions, images and feedback provided on each question at the end of the tool to facilitate students' future learning pattern.

"Challenges" covered the codes – "challenging in general," "to incorporate in the curriculum," "prosection images difficulty," "radiology images difficulty," and "clinical context too complex." These highlighted the difficulty that students found to comprehend the content and to interpret anatomical and radiological images.

Test applicability/quality. Students valued the inclusion of clinical/applied information in the test. They commented on how such questions made them think about multiple levels of topographical, functional, and applied anatomy. They showed an understanding of the validity of both anatomical and radiological images in making concepts and application in clinical settings, respectively. Moreover, they found the formative feedback provided on each question useful.

"Hi, I thought the quiz was excellent. Very clinically relevant and an excellent revision tool. I had to think back to all my anatomy knowledge! For the exams at my medical school, these types of questions match the kind of questions that come up in exams so for me personally, it was an excellent revision tool. I would definitely use this type of resource if it was made available. The questions were ideal in length (not too wordy) and very clear. I enjoyed this quiz."

"Really an excellent test. What made it better than most examinations of medical knowledge was that sort of 'extra step' you had in many questions. For example, instead of asking simply what innervated the upper larynx, you asked what might cause a cough reflex there. We learn so much of our course through text that when I get to a question about, say, the lumbar puncture layers, I'm made to look deep into my knowledge of the structure and use many of those text-based facts I know to answer the

single question. Standard examination questions often do not do this and rather rely on us to just remember single-sentence obscurities from lectures to assess our depth of knowledge. Thank you very much and I hope my results are useful!"

Challenges. Students found it challenging to answer an anatomical question formatted in a clinical/applied scenario, interpret anatomical and radiological images. Some students also commented on the perceived mismatch between their style of learning and curriculum and this test.

"This experience has highlighted how little anatomy is taught at my medical school and how when presented with an image of a cadaver we are stumped. Anatomy at my medical school is primarily taught with coloured images, models and living anatomy. When the colour is taken away and we are presented with surgical or cadaveric dissection images we are left at a loss as to how to identify structures."

"The questions provided a good practical application of anatomy. However, during our teaching, the practical side has not been emphasised as much oppose to learning the theory hence making the 'jump' was something quite difficult – especially as we are taught with some radiology images but not many. This left me being unable to work out which side of the body was shown or which ligament etc. although I knew the knowledge."

Very difficult to understand 3D structures from pictures of prosections. Questions were good and challenging but at my level of study, it felt like a bit too much emphasis on the precise clinical manifestation. Questions on clinical manifestations are important but for a 2nd year the basics being tested too would be good, you might be overestimating my abilities!"

DISCUSSION

The findings are discussed in the light of literature, and it is an incremental contribution to the field of the effect of the images in online anatomy assessments.

The mean performance of the students across the six schools was variable. This variability is likely due to the uneven distribution of the sample size per school and/or the variance in the competence of the participants. Out of 174 students, 78 were low

performing and 96 were high performing, and there were more high performing students from School 3 as compared to the other five schools. Therefore, the mean scores do not reflect the similarity of the level of competence across the schools. However, based on academic leads reviews, it was assumed that the test was a reliable measure to test their anatomy knowledge at the end of year 2.

Students generally scored higher on questions with images. This is in keeping with the assumptions of the CTML that people learn better from a combination of words and images (Mayer, 2005b) and emphasizes the role of images in simplifying accompanying text (Levie and Lentz, 1982; Winn, 1989; Peeck, 1993; Carney and Levin, 2002). However, although images are known to facilitate learning, the literature shows a variable effect on individual questions (Crisp and Sweiry, 2006; Vorstenbosch et al., 2013), the involvement of different cognitive processes (Vorstenbosch et al., 2014), no effect of visual resources (Buzzard and Bandaranayake, 1991; Khalil et al., 2005; Inuwa et al., 2011, 2012; Notebaert, 2017; Bahlmann, 2018), and increase in question difficulty (Berends and van Lieshout, 2009). In fact, Vorstenbosch et al. (2013) and Hunt (1978) demonstrated that students found image-based questions easier than text-based ones. In contrast, Holland et al. (2015) have shown no significant difference in item discrimination or difficulty with and without the inclusion of an image. The improved performance of students on questions with images in this study may thus imply their ability to interpret these images successfully. However, it could also suggest easy retrieval of pre-existing knowledge because of previous exposure to similar images during learning. On the other hand, it may also suggest the students' lack of adequately developed schemas to effectively interpret the clinical concepts without the aid of an image, particularly where they performed poorly on questions without images (Sweller, 1994; Regehr and Norman, 1996).

However, the type of image (anatomical or radiological) utilized, i.e., its "surface structure," had no significant influence on the students' mean scores. This is in keeping with the work of Khalil et al., (2005), Schubert et al., (2009), and Inuwa et al., (2011, 2012) that showed no significant differences in mean scores on image questions. Unlike the present study, these previous studies were based on questions for the immediate recall of anatomical information. Nevertheless, Crisp and Sweiry (2006) showed that differences in the images significantly affected scores of one question and had smaller effects on the others. In Berends and van Lieshout's (2009) study, the presence of images increased item difficulty and slowed down the speed at which students were able to process information.

In contrast, analysis of the influence of question subtypes and deep structures (bones and soft tissue) in anatomical and radiological images showed a highly significant adverse effect on the students' scores. This may be the result of an inability to process the relatively more layers of information required to answer image questions containing soft tissues compared to those containing bones. Layers of information refer to the details that images with soft tissues often have because of inter-related structures such as muscles, nerves, arteries, and veins. Furthermore, students classically start to build anatomical knowledge from bones outward, and this is reflected in anatomy textbooks, atlases and our ways of using osteology as a scaffolding for teaching anatomy. Therefore, reiterative reviews of the same information over time would explain their better performance on bone questions.

It is clear that anatomical and radiological images usually focus on a discrete body region and do not usually depict the entirety of the body. This separation of the part from the whole in the images, therefore, demands the elicitation of multiple cognitive processes to identify the part, determine its location and orientation in the whole, and simultaneously interpret the relationship of neighboring structures. This gets more complex in images with structures of different densities compared to images with only bones. The finding that the perceptual surface and deep semantic structure of nonhomogeneous images significantly influences students' performance, thus reinforces the application of the Alternative Multimedia Learning Theory to anatomical and radiological assessments. In addition to the use of valid and authentic images in such assessments, it is therefore essential to take the perceptual surface and deep semantic structure of these images into consideration.

Furthermore, the unprompted comments from the participants highlighted their perceptions which helped to understand and analyze the data critically and their performance on these clinically oriented questions in an online platform. The qualitative data showed that students had a clear preference for being tested in a clinical/applied context. This highlights the importance of assessments in the "knows how" category of Miller's pyramid (Miller, 1990), and the application of knowledge in the modified Bloom's taxonomy (Bloom, 1956). The influence of the authenticity of anatomical images in building schemas and the validity of radiological images in clinical settings was evident in the students' emphasis of the importance of using pre-existing knowledge to orient and decipher these images, and to develop and utilize the appropriate schemas in these contexts (Schnotz, 2002). For some, image questions were useful to visualize a clinical or applied scenario. However, for others, interpreting images required extra-cognitive effort (Schnotz and Bannert, 2003) demonstrating how images could interfere or support the orchestration between internal and external representations. The challenge of interpreting anatomical images suggests that the cognitive transition of making sense of three-dimensional structures in two-dimensional images is not a smooth, intuitive process. Also, the inclusion of color in drawings or illustrations of anatomical structures could eventually interfere rather than facilitate the process of interpreting anatomical and radiological images. This matter requires further research.

Hence, the findings suggest that the students' performance on clinically oriented anatomy questions with and without images is dependent on an intricate network of factors; including perceptual surface structure, deep semantic structure, orchestration between existing schemas with external representation (text and images) and question difficulty.

Limitations of the Study

This was originally designed to be a multi-institutional study to compare the teaching style and assessment performance of the six medical schools; however, most of the data being from one medical school limits the study's multi-institutional aspect. Further, the uneven distribution of numbers of participating students from the six medical schools meant that comparisons of any effects of the teaching resources and assessments practices of individual schools on the performance of the students on the different question types were not statistically possible. Based on academic leads review of the test, it was assumed that the test was a reliable tool for assessing their anatomy knowledge at the end of year 2. The study was limited to a single

time-limited simultaneous contact window with the six medical schools. As such, it was not possible to randomize the participants and compute baseline performance data of students from the participating schools. However, the normality tests showed that the data were normally distributed.

CONCLUSION

The analysis showed that students' performance was significantly higher on clinically oriented anatomy questions with images compared to questions without images. No significant difference in performance was seen between the questions with an anatomical or radiological image. Further analysis indicated that students performed significantly higher on questions referring to bones compared to those referring to soft tissues, which suggests that the deep semantic structure of an image has a significant impact on students' performance.

Along with this, students valued the inclusion of clinical information and images in the test. Also, they appreciated the feedback provided on each question for their future learning. The data also highlighted the challenges of interpreting various types of images used in anatomy.

The principal implication of the findings is that images impact students' performance in applied anatomy SBA assessments, and teachers and examiners ought to take this into account in designing these assessments and interpreting the results. Moreover, the deep semantic structure of the images has shown to play a significant role; therefore, questions referring to bones and soft tissues should be one of the criteria for blueprinting, and the analysis of results should take the students' performance on these supplementary and nonhomogenous images into consideration.

ACKNOWLEDGMENTS

NOTES ON CONTRIBUTORS

MANDEEP GILL SAGOO, Ed.D., M.Sc. (Anat.), P.G.C.E., F.H.E.A., is a lecturer in anatomy in the Department of Anatomy, School of Centre for Education at King's College London in London, United Kingdom. She is a lead for M.B.B.S. year 1 assessments and her research is in understanding assessments with its relation to cognitive psychology of multimedia learning, and spatial ability.

MARC A.T.M. VORSTENBOSCH, Ph.D., is an associate professor in the Department of Anatomy at Radboud University Nijmegen Medical Centre, Radboud, The Netherlands. He teaches anatomy of the head and neck and the locomotor system and his research is in anatomy education, spatial learning, and assessment.

PETER J. BAZIRA, M.B. Ch.B., M.Sc., Ed.D., S.F.H.E.A., is Professor of Clinical Anatomy and Medical Education

and the Director of the Centre for Anatomical and Human Sciences at Hull York Medical School, York, United Kingdom. His research is in anatomy education and authenticity in learning in the context of medical and health professions education.

HAROLD ELLIS, C.B.E., F.R.C.S., is a retired surgeon and professor in the Department of Anatomy, School of Biomedical Education at King's College London in London, United Kingdom. He is one of the most notable British surgeons, author of many anatomy and surgery books, and an inspirational teacher.

MARIA KAMBOURI, Ph.D., is a senior lecturer in research methods in the Department of Psychology and Human Development at Institute of Education, University College London, London, United Kingdom. Her research is in adult learning and innovative pedagogies, especially in information and communication technology.

CHARLIE OWEN, Ph.D., is a principal research associate in Thomas Coram Research Unit at Institute of Education, University College London in London, United Kingdom. His research is in social research methods, particularly in the secondary analysis of official statistics.

LITERATURE CITED

Ahmed K, Rowland S, Patel V, Khan RS, Ashrafian H, Davies DC, Darzi A, Athanasiou T, Paraskeva PA. 2010. Is the structure of anatomy curriculum adequate for safe medical practice? Surgeon 8:318–324.

Ainsworth S. 1999. The functions of multiple representations. Comput Educ 33:131–152.

Baddeley AD. 1992. Working memory. Science 255:556–559.

Bahlmann O. 2018. Illustrated versus non-illustrated anatomical test items in anatomy course tests and German Medical Licensing examinations (M1). GMS J Med Educ 35:Doc25.

Berends IE, van Lieshout EC. 2009. The effect of illustrations in arithmetic problem-solving: Effects of increased cognitive load. Learn Instruct 19:345–353.

BERA. 2018. British Educational Research Association. Ethical Guidelines for Educational Research. 4th Ed. London, UK: British Educational Research Association. 41 p. URL: https://www.bera.ac.uk/wp-content/uploads/2018/06/BERA-Ethical-Guidelines-for-Educational-Research_4thEdn_2018.pdf [accessed 6 January 2020].

Biedermann I. 1981. On the semantics of a glance at a scene. In: Kubovy M, Pomerantz JR (Editors). Perceptual Organization. 1st Ed. Hillsdale, NJ: Lawrence Erlbaum Associate, Inc. p 213–254.

Biggs JB, Tang C. 2011. Teaching for Quality Learning at University. 4th Ed. Maidenhead, UK: Open University Press. 480 p.

Bloom BS (Editor). 1956. Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain. 1st Ed. New York, NY: David McKay Co. 201 p.

Boshuizen HP, Schmidt HG. 1992. On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. Cognit Sci 16:153–184.

Brenner E, Chirculescu AR, Reblet C, Smith C. 2015. Assessment in anatomy. Eur J Anat 19:105–124.

Buzzard AJ, Bandaranayake RC. 1991. Comparison of the performance of visual and verbal multiple-choice questions. Aust N Z J Surg 61:614–618.

Carney R, Levin J. 2002. Pictorial illustrations still improve students' learning from text. Educ Psychol Rev 14:5–26.

Case SM, Swanson DB. 2002. Constructing Written Test Questions for the Basic and Clinical Sciences. 3rd Ed. Philadelphia, PA: National Board of Medical Examiners (NBME). 180 p.

Chandler P, Sweller J. 1991. Cognitive load theory and the format of instruction. Cognit Instruct 8:293–332.

Crisp VM, Sweiry E. 2006. Can a picture ruin a thousand words? The effects of visual resources in exam questions. Educ Res 48:139–154.

Cronbach LJ. 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16:297–333.

Dettmer S, Schmiedl A, Meyer S, Giesemann A, Pabst R, Weidemann J, Wacker FK, Kirchhoff T. 2013. Radiological anatomy - Evaluation of integrative education in radiology. Rofo 185:838–843.

Doane DP, Seward LE. 2011. Measuring skewness. J Stat Educ 19:2.

Engelhardt PV. 2009. An introduction to classical test theory as applied to conceptual multiple-choice tests. In: Henderson C, Harper KA (Editors). Getting Started in PER. Volume 2. 1st Ed. College Park, MD: American Association of Physics Teachers. p 1–40.

Ginns P. 2005. Meta-analysis of the modality effect. Learn Instruct 15:313–331.

GMC. 2009. General Medical Council. Tomorrow's Doctors. Outcomes and Standards, for Undergraduate Medical Education. Regulating Doctors, Ensuring Good Medical Practice. 3rd Ed. London, UK: General Medical Council. 104 p.

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

Haladyna TM, Rodriguez MC. 2013. Developing and Validating Test Items. 1st Ed. New York, NY: Routledge. 454 p.

Holland J, O'Sullivan R, Arnett R. 2015. 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 15:184.

Hunt DR. 1978. Illustrated multiple choice examinations. Med Educ 12:417–420.

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.

King N, Horrocks C. 2010. Interviews in Qualitative Research. 1st Ed. London, UK: Sage Publications, Ltd. 256 p.

Landis JR, Koch GG. 1977. The measurement of observer agreement for categorical data. Biometrics 33:159–174.

Levie WH, Lentz R. 1982. Effects of text illustrations: A review of research. Educ Comm Tech J 30:195–232.

Louw G, Eizenberg N, Carmichael SW. 2009. The place of anatomy in medical education: AMEE Guide no 41. Med Teach 31:373–386.

Mayer RE. 2005a. Cognitive theory of multimedia learning. In: Mayer RE (Editor). The Cambridge Handbook of Multimedia Learning. 1st Ed. Cambridge, UK: Cambridge University Press. p 31–48.

Mayer RE. 2005b. Introduction to multimedia learning. In: Mayer RE (Editor). The Cambridge Handbook of Multimedia Learning. 1st Ed. Cambridge, UK: Cambridge University Press. p 1–26.

Mayer RE. 2009. Multimedia Learning. 2nd Ed. New York, NY: Cambridge University Press. 320 p.

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

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

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

MSC. 2020. Medical Schools Council. Entry requirements. Medical Schools Council, London, UK. URL: https://www.medschools.ac.uk/studying-medicine/applications/entry-requirements [accessed 6 January 2020].

Notebaert AJ. 2017. The effect of images on item statistics in multiple choice anatomy examinations. Anat Sci Educ 10:68–78.

Older J. 2004. Anatomy: A must for teaching the next generation. Surgeon 2:79–90.

Orsbon CP, Kaiser RS, Ross CF. 2014. Physician opinions about an anatomy core curriculum: A case for medical imaging and vertical integration. Anat Sci Educ 7:251–261.

Paivio A. 1986. Mental Representations: A Dual Coding Approach. 1st Ed. New York, NY: Oxford University Press. 332 p.

Peeck J. 1993. Increasing picture effects in learning from illustrated text. Learn Instruct 3:227–238.

Phillips AW, Smith SG, Straus CM. 2013. The role of radiology in preclinical anatomy: A critical review of the past, present, and future. Acad Radiol 20:297. e1–304.e1.

Regehr G, Norman GR. 1996. Issues in cognitive psychology: Implications for professional education. Acad Med 71:988–1001.

Robson C. 2011. Real World Research. 3rd Ed. Chichester, UK: John Wiley & Sons Ltd. 608 p.

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 Radiol Anat 33:263–269.

Samarasekera DD, Gopalakirshnakone P, Gwee MC. 2015. Assessing anatomy as a basic medical science. In: Chan LK, Pawlina W (Editors). Teaching Anatomy: A Practical Guide. 1st Ed. New York, NY: Springer International Publishing. p 279–289.

Schmidt HG, Rikers RM. 2007. How expertise develops in medicine: Knowledge encapsulation and illness script formation. Med Educ 41:1133–1139.

Schnotz W. 2002. Towards an integrated view of learning from text and visual displays. Educ Psychol Rev 14:101–120.

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

Schnotz W, Bannert M. 2003. Construction and interference in learning from multiple representation. Learn Instruct 13:141–156.

Schubert S, Schnabel KP, Winkelmann A. 2009. Assessment of spatial anatomical knowledge with a 'three-dimensional multiple-choice test' (3D-MC). Med Teach 31:e13–e17.

Schuwirth LW, van der Vleuten CP. 2004a. Changing education, changing assessment, changing research? Med Educ 38:805–812.

Schuwirth LW, van der Vleuten CP. 2004b. Different written assessment methods: What can be said about their strengths and weaknesses? Med Educ 38:974–979.

Schuwirth LW, van der Vleuten CP. 2011. Programmatic assessment: From assessment of learning to assessment for learning. Med Teach 33:478–485.

Shapiro SS, Wilk MB. 1965. An analysis of variance test for normality. Biometrika 52:591–611.

Sugand K, Abrahams P, Khurana A. 2010. The anatomy of anatomy: A review for its modernization. Anat Sci Educ 3:83–93.

Sweller J. 1994. Cognitive load theory, learning difficulty, and instructional design. Learn Instruct 4:295–312.

Van der Vleuten CP, Schuwirth LW. 2005. Assessing professional competence: From methods to programmes. Med Educ 39:309–317.

Vorstenbosch MA, Bouter ST, van den Hurk MM, Kooloos JG, Bolhuis SM, Laan RF. 2014. Exploring the validity of assessment in anatomy: Do images influence cognitive processes used in answering extended matching questions? Anat Sci Educ 7:107–116.

Vorstenbosch MA, Klaassen TP, Kooloos JG, Bolhuis SM, Laan RF. 2013. Do images influence assessment in anatomy? Exploring the effect of images on item difficulty and item discrimination. Anat Sci Educ 6:29–41.

Vorstenbosch MA, Kooloos JG, Bolhuis SM, Laan RF. 2016. An investigation of anatomical competence in junior medical doctors. Anat Sci Educ 9:8–17.

Wilson MH, Wise D, Davies G, Lockey D. 2012. Emergency burr holes: "How to do it". Scand J Trauma Resusc Emerg Med 20:24.

Winn W. 1989. The design and use of instructional graphics. In: Mandl H, Levin JR (Editors). Knowledge Acquisition from Text and Pictures. 1st Ed. Amsterdam, The Netherlands: Elsevier Science Publishers B.V. p 125–144.

Witteman MJ, Segers E. 2010. The modality effect tested in children in a user paced multimedia environment. J Comput Assist Learn 26:132–142.

Wood T, Cole G, Lee C. 2004. Developing Multiple Choice Questions for RCPSC Certification Examinations. 1st Ed. Ottawa, Ontario, Canada: Office of Education, The Royal College of Physicians and Surgeons of Canada. 10 p.