



# Volumetric CT-images improve testing of radiological image interpretation skills



Cécile J. Ravesloot<sup>a,\*</sup>, Marieke F. van der Schaaf<sup>b</sup>, Jan P.J. van Schaik<sup>a</sup>, Olle Th.J. ten Cate<sup>c</sup>, Anouk van der Gijp<sup>a</sup>, Christian P. Mol<sup>d</sup>, Koen L. Vincken<sup>d</sup>

<sup>a</sup> Radiology Department at University Medical Center Utrecht, Heidelberglaan 100, 3508 GA Utrecht, Room E01.132, The Netherlands

<sup>b</sup> Department of Pedagogical and Educational Sciences at Utrecht University, Heidelberglaan 1, 3584 CS Utrecht, The Netherlands

<sup>c</sup> Center for Research and Development of Education at University Medical Center Utrecht, Heidelberglaan 100, 3508 GA Utrecht, The Netherlands

<sup>d</sup> Image Sciences Institute at University Medical Center Utrecht, Heidelberglaan 100, 3508 GA Utrecht, The Netherlands

## ARTICLE INFO

### Article history:

Received 4 June 2014

Received in revised form

11 December 2014

Accepted 12 December 2014

### Keywords:

Education

Radiology test

Test quality

Volumetric images

## ABSTRACT

**Rationale and objectives:** Current radiology practice increasingly involves interpretation of volumetric data sets. In contrast, most radiology tests still contain only 2D images. We introduced a new testing tool that allows for stack viewing of volumetric images in our undergraduate radiology program. We hypothesized that tests with volumetric CT-images enhance test quality, in comparison with traditional completely 2D image-based tests, because they might better reflect required skills for clinical practice.

**Materials and methods:** Two groups of medical students ( $n = 139$ ;  $n = 143$ ), trained with 2D and volumetric CT-images, took a digital radiology test in two versions (A and B), each containing both 2D and volumetric CT-image questions. In a questionnaire, they were asked to comment on the representativeness for clinical practice, difficulty and user-friendliness of the test questions and testing program. Students' test scores and reliabilities, measured with Cronbach's alpha, of 2D and volumetric CT-image tests were compared.

**Results:** Estimated reliabilities (Cronbach's alphas) were higher for volumetric CT-image scores (version A: .51 and version B: .54), than for 2D CT-image scores (version A: .24 and version B: .37). Participants found volumetric CT-image tests more representative of clinical practice, and considered them to be less difficult than volumetric CT-image questions. However, in one version (A), volumetric CT-image scores ( $M 80.9$ ,  $SD 14.8$ ) were significantly lower than 2D CT-image scores ( $M 88.4$ ,  $SD 10.4$ ) ( $p < .001$ ). The volumetric CT-image testing program was considered user-friendly.

**Conclusion:** This study shows that volumetric image questions can be successfully integrated in students' radiology testing. Results suggests that the inclusion of volumetric CT-images might improve the quality of radiology tests by positively impacting perceived representativeness for clinical practice and increasing reliability of the test.

© 2015 Elsevier Ireland Ltd. All rights reserved.

## 1. Introduction

Advanced technological developments have drastically changed radiological practice, especially cross-sectional imaging modalities which changed from viewing series of single cross-sections printed next to each other (tile viewing), to scrolling through several hundreds of images (stack viewing), with the possibility to

adjust viewing directions and contrast settings [1,2]. This change from tile to stack viewing demands different and new skills of radiologists and other medical specialists involved in interpreting large volumetric datasets [1,3].

Recent research shows that integrating volumetric datasets (images) in radiology education is an effective way to increase image interpretation skills of medical students [4]. In contrast, the testing of radiological image interpretation skills in undergraduate and postgraduate medical education is still predominantly based on tests that exclusively contain questions concerning one or several cross-sections of a CT or MRI scan (2D image test). Participants of 2D image tests cannot scroll through or manipulate the images (cross-sections), including changing viewing direction (multiplanar reformatting), zooming, and adapting tissue contrast. As a result, image manipulation skills are not tested, while this is

\* Corresponding author. Tel.: +31 887550355/887556688; fax: +31 302501098.

E-mail addresses: [C.J.Ravesloot@umcutrecht.nl](mailto:C.J.Ravesloot@umcutrecht.nl) (C.J. Ravesloot),

[M.F.vanderSchaaf@uu.nl](mailto:M.F.vanderSchaaf@uu.nl) (M.F. van der Schaaf), [J.P.J.vanSchaik@umcutrecht.nl](mailto:J.P.J.vanSchaik@umcutrecht.nl)

(J.P.J. van Schaik), [T.J.tenCate@umcutrecht.nl](mailto:T.J.tenCate@umcutrecht.nl) (O.Th.J. ten Cate),

[A.vanderGijp-2@umcutrecht.nl](mailto:A.vanderGijp-2@umcutrecht.nl) (A. van der Gijp), [C.Mol@umcutrecht.nl](mailto:C.Mol@umcutrecht.nl) (C.P. Mol),

[K.Vincken@umcutrecht.nl](mailto:K.Vincken@umcutrecht.nl) (K.L. Vincken).

considered to be important in image interpretation [5]. Moreover, in 2D image tests the slice of interest is provided, and participants do not need to scroll to search for abnormalities. This raises questions about the quality of these 2D image tests for assessing currently needed radiological image interpretation skills. Volumetric images, i.e. volumetric CT or MR datasets allowing for stacking viewing, might increase test quality.

Test quality requirements include reliability and validity [6,7]. Reliability focuses on the accuracy of the results, i.e. on the precision and reproducibility of the test scores. Validity means that the test is properly testing what is intended to be tested, and that test scores correctly reflect intended skills or knowledge. Authenticity can add to validity, as authentic tests reflect clinical practice and are likely to capture clinically relevant skills and knowledge. The validity of radiological image interpretation tests exclusively based on 2D images is questionable, since they no longer reflect authentic practice. Other important test quality requirements are fairness and usability [6,7]. Fairness implies that inferences made from the test results should be transparent and legitimate. For example, the difficulty of a test should be in accordance with the expected level of expertise of the participants and with the difficulty of the educational program. Otherwise, students who acquired sufficient knowledge and skills could fail their exams, or vice versa. To achieve optimal usability, the costs and complexity of the test should be low and testing tools should be intuitive and user-friendly.

This study aims to compare the quality of volumetric versus 2D CT-image questions for testing radiological image interpretation skills. Indications of reliability, authenticity, fairness, difficulty and usability of 2D and volumetric CT-image questions were studied to evaluate test quality. We hypothesized that volumetric CT-images positively impact test quality, while keeping the testing program user-friendly.

## 2. Materials and methods

### 2.1. Context of the study

At the University Medical Center Utrecht (The Netherlands), second year undergraduate medical students take a radiology test as part of their radiology educational program. Learning objectives are image interpretation of radiological anatomy and prevalent illnesses. Radiological CT-anatomy is taught using both volumetric and 2D CT-images. Before 2010 CT-anatomy questions in the test

exclusively contained 2D images of CT-scans (tile viewing). In order to improve the test quality, part of the 2D CT-image questions were replaced by volumetric CT-image questions in 2010 using a novel digital testing program called “VQuest” [8]. VQuest was designed to test volumetric image interpretation skills, and allows for stack viewing, including view and contrast changing (Fig. 1 and video).

### 2.2. Study design

In April 2010, a radiology test in two versions (A and B) was evaluated to study quality aspects of volumetric CT-image questions compared to the 2D CT-image questions. Reliability was evaluated by analysing the internal consistency of 2D and volumetric test scores. An indication of fairness was obtained by comparing test scores of 2D and volumetric image questions and by investigating the participants' opinion about the difficulty differences of both 2D and volumetric image questions. Immediately after the test, participants were asked to fill out a questionnaire including questions on user-friendliness of the testing program (as indication for usability) and perceived representativeness for clinical practice (authenticity) of 2D and volumetric image questions on paper. Participants were informed that the questionnaire was voluntary and used for evaluation of the new volumetric image questions and testing program. All test results and questionnaire responses were anonymized. Because analyses of test results and questionnaires were part of the regular quality control cycle of the educational program and all data were anonymized, no signed informed consent was obtained. The study was approved by the Ethical Review Board of the Netherlands Association for Medical Education.

### 2.3. Participants

The test was taken by 282 second-year medical students. All students followed the same educational program. Because of the limited availability of computers the participants were divided in two groups and tested at different moments in one week. Both groups took a different version of the test (version A or B) to prevent sharing of answers. Prior to the test all participants were invited to attend a meeting to practise with volumetric CT-image questions based on stack viewing in the testing program VQuest. Almost all participants attended this meeting. 277 participants filled out the questionnaire (response rate was 98%).



**Fig. 1.** Student interface of VQuest (2013), digital testing tool for volumetric images, showing an example of a question in which the student is asked to mark an anatomical structure.

## 2.4. Radiology test

The objectives of the test were basic radiological knowledge and image interpretation skills on prevalent illnesses and radiological anatomy. The test contained 50 questions in total and lasted 1.5 h. Both versions of the test contained ten volumetric CT-image questions and eleven 2D CT-image questions, all on CT anatomy (abdomen, chest or brain). In case of a volumetric CT-image question, participants were asked to label a specific anatomical structure (Fig. 1). The 2D image questions were designed as long-menu multiple-choice questions (including about 25 options): an anatomical structure was marked in a CT-slice, and participants were asked to identify it and choose the correct option from the list. The volumetric image questions were applied in the digital testing program VQuest and were viewed in stack mode (see video). 2D CT-image questions were assessed using Testvision®, a digital testing program which does not allow stack viewing. Participants were not able to change viewing direction or tissue contrast of either 2D or volumetric CT-images. A team of radiology residents and an experienced radiologist who were involved in the radiology teaching of medical students constructed the test questions conform the learning objectives of the educational program. Four 2D CT-image questions in both versions of the test were not used in prior exams. All other 2D image questions had been used previously and were considered to be good quality questions, around 80% of them had an  $r$ -value (discrimination value) above .15. Anatomical structures for the newly developed 20 volumetric CT-anatomy test questions were selected from the learning objectives list for second year medical students.

## 2.5. Questionnaire

The questionnaire was based on a survey that is regularly administered for the evaluation of tests of undergraduate medical education at the University Medical Center of Utrecht and was developed by the Center for Research and Development of Medical Education. Specific questionnaire items for the evaluation of the 2D and volumetric CT-image questions were designed by the examiners of the radiology test in consultation with an evaluation employee of the Center for Research and Development of Education at and an associate professor at the Department of Pedagogical and Educational Sciences of Utrecht University, and added to the questionnaire. Fifteen items were used to evaluate perceived representativeness for clinical practice (1 item), difficulty (5 items) and user-friendliness (9 items) of the volumetric image questions compared to traditional 2D image questions and are shown in Table 2. Participants could rate their agreement with each item on a five-point Likert scale, in which 1 is “not agree”, and 5 is “completely agree”. Because there were several items measuring the same subject, a reliability scale (Cronbach's alpha) could be obtained indicating consistency among questions, i.e. all participants' responses on these questions result in the same score [9,10]. Cronbach's alpha is a correlation coefficient that reaches 1.0 when all responses on questions of the same scale are exactly equal. Scales with a Cronbach's alpha coefficient of .8 or higher are considered reliable [10]. Both scales measuring perceived user-friendliness of the testing program and perceived difficulty of volumetric versus 2D image questions had good estimated reliabilities of Cronbach's alpha .81.

## 2.6. Data analysis

### 2.6.1. Reliability of the radiology test

Reliability estimates for test scores on volumetric and 2D image questions were computed with Cronbach's alpha. Because Cronbach's alpha is highly dependent on the number of questions in

**Table 1**

Radiology test results: estimated reliability (Cronbach's  $\alpha$ ), means and standard deviations of percentage scores of 2D and volumetric image questions.

Test version (and $N$ participants)	Version A (139)	Version B (143)
2D image questions		
$\alpha$ ( $k$ )	.24 (11)	.37 (11)
Mean percentage score (SD)	88.4 (10.4)	82.3 (11.7)
Volumetric image questions		
$\alpha$ ( $k$ )	.51 (10)	.54 (10)
Mean percentage score (SD)	80.9 (14.8)	79.9 (15.5)

$k$  is the number of questions.

the test, tests with fewer questions, as in our study (10 volumetric image questions and 11 2D image questions) will have low alphas [10]. To estimate the number of 2D and volumetric image questions needed for a Cronbach's alpha of .80 the Spearman Brown Formula was used: number of questions =  $(k \times .80) \times (\alpha - 1) / (\alpha \times (.80 - 1))$ , where  $\alpha$  stands for Cronbach's alpha of the current test and  $k$  is the number of questions in the current test (for 2D image test  $k = 11$  and for volumetric image test  $k = 10$ ) [9].

### 2.6.2. Difficulty of the radiology test

Indications of objective difficulty differences between 2D and volumetric CT-image questions were obtained by comparing means and standard deviations of scores for 2D and volumetric CT-image questions per version using paired  $t$ -tests. Further  $p$ -values of 2D and volumetric CT-image questions were compared. The  $p$ -value is the proportion of correct answers for a test question, so the higher the  $p$ -value, the easier the question.

For each participant a score for perceived difficulty was calculated by averaging the responses on the three questions concerning the perceived difficulty scale (light blue shaded items 2–4 in Table 2). Perceived difficulty was evaluated by calculating the mean score and standard deviation of all responses on the perceived difficulty scale. This resulted in a mean score between 1 and 5, where a score higher than 2.5 indicated that participants considered volumetric CT-image questions easier than 2D CT-image questions. Added to these calculations, differences between mean scores on two paired items on 2D and volumetric CT-image question difficulty (see Table 2, question numbers 5 and 6) were tested using paired  $t$ -tests, after assumption checks.

### 2.6.3. Perceived representativeness of clinical practice

Perceived representativeness was evaluated by calculating the mean and standard deviation of the responses on questionnaire item number 1, resulting in a score between 1 and 5. A score higher than 2.5 indicated that participants considered volumetric CT-image questions more representative for clinical practice than 2D CT-image questions.

### 2.6.4. User-friendliness of the testing program

To get an indication of the user-friendliness of the volumetric image testing program the mean scale score of the responses on items measuring user-friendliness was calculated (see Table 2 item number 7–12). This resulted in a mean score between 1 and 5, where a score higher than 2.5 indicated that participants considered the volumetric CT-image testing program user-friendly.

## 3. Results

### 3.1. Reliability of volumetric image questions

Cronbach's alpha for scores on volumetric CT-image questions was higher than for scores on 2D CT-image questions (see Table 1), indicating a higher reliability. The Spearman Brown Formula

**Table 2**

Questionnaire results: means (M) and standard deviations (SD) of responses on questionnaire items on perceived representativeness of clinical practice, difficulty and user-friendliness of volumetric and 2D image questions. All questions are translated from Dutch to English.

Item number		M	SD	Number of responses
Perceived representativeness				
Scale: 1–5 (“completely disagree” to “completely agree”)				
1	Volumetric image questions reflect clinical practice better than 2D image items.	4.1	0.8	267
Perceived difficulty				
Scale: 1–5 (“completely disagree” to “completely agree”)				
2	Mentally representing a 3D image of anatomical structures out of a volumetric image is easier than out of a 2D image.	4.4	0.7	274
3	Recognizing anatomical structures in a volumetric image is easier than in a 2D image.	4.4	0.8	273
4	Stack viewing increases my ability to mentally represent relations between anatomical structures.	4.5	0.6	273
5	Indicating anatomical structures in a 2D image is easy.*	2.9	0.9	268
6	Indicating anatomical structures in a volumetric image is easy.*	3.8	0.9	262
User-friendliness				
Scale: 1–5 (“completely disagree” to “completely agree”)				
7	The user-friendliness of ..... is good.	4.2	0.7	275
8	The instruction manual of .... is clear.	4.2	0.7	274
9	The interface of .... is clear.	4.1	0.7	275
10	The questions in .... were comprehensible.	4.1	0.7	272
11	The image quality in .... was high.	4.0	0.9	274
12	The time to complete the test was sufficient.	4.5	0.6	276
13	Navigating between questions is fast.	4.1	0.7	273
14	The unsure box† is useful.	4.0	1.0	267
15	The progress bar‡ is useful.	3.9	0.8	263

\* Significantly different for 2D and volumetric image questions at  $p < .001$  using paired  $t$ -test.

† Students can mark questions in case of uncertainty by checking the “unsure box”. In this way they can easily recognize them and reopen them to make adjustments to their answers.

‡ The progress bar shows how many questions are already answered during the test.

Perceived difficulty scale is calculated by averaging the scores on light blue shaded questions (no 2–4) for each participant.

predicts that fewer volumetric image questions are needed to achieve a desirable reliability (.80 or higher) compared to 2D CT-image questions: 38 and 34 volumetric CT-image questions compared to 139 and 75 2D CT-image questions, for version A and B respectively.

### 3.2. Difficulty

The mean of responses on the scale items measuring volumetric image question difficulty compared to 2D CT-image question difficulty was 4.5 (scale 1–5) (SD=0.6;  $n=268$ ), see Table 2. This indicates that participants considered volumetric image questions less difficult than 2D CT-image questions. This is congruent with the significant difference in mean score of responses on the difficulty items 5 and 6 ( $t(256)=11.89$ ,  $p<.001$ ; eta squared=0.36). In contrast, the mean percentage radiology test scores on volumetric CT-image questions were lower than mean percentage scores on 2D CT-image questions (see Table 1). This difference was significant only for version A of the test ( $t(138)=5.50$ ,  $p<.001$ , eta squared 0.18). This difference in difficulty between 2D and volumetric CT-image questions is also shown by the difference in  $p$ -values of both question types.  $p$ -Values of the majority of 2D CT-image questions were higher than .80 (ten 2D CT-image questions in version A and nine in version B). In version B there was one 2D CT-image question with a  $p$ -value of .24 (24% of the participants gave the correct

answer), which appeared to be a relatively difficult question compared to the other 2D CT-image questions in the test.

### 3.3. Perceived representativeness of clinical practice

Mean score for questionnaire responses on perceived representativeness was 4.1 (SD .84) on a five point Likert scale, see Table 2, indicating that participants considered volumetric CT-image questions more representative for clinical practice than 2D CT-image questions.

### 3.4. User-friendliness of volumetric image testing program

The mean of responses on the user-friendliness scale of the volumetric CT-image testing program was 4.1 (scale 1–5) (SD=.47;  $n=277$ ). See Table 2 for means of responses on individual items (no 7–15).

## 4. Discussion

The estimated reliability of volumetric image question scores was higher than for 2D CT-image question scores in both test versions. There were more questions needed for a desirable reliability when using 2D CT-images compared to volumetric images. This indicates that volumetric image question scores are



more accurate and more reproducible compared to 2D CT-image question scores. This might be due to a greater dispersion in scores for volumetric image questions compared to 2D image questions (see Table 1). Possibly, participants need to master more or other kinds of knowledge and skills to complete a volumetric CT-image test. This can lead to a greater dispersion of scores, and better discriminating power between high and low performers on the image interpretation task. However these results should be interpreted with care, because of the relative low number of 2D and volumetric image questions in the tests. This might have influenced test score reliabilities, which were fairly low. Therefore, evidence from larger empirical prospective research is needed for confirmation and reproducibility of the results.

In version A of the test, the scores on volumetric CT-image questions were significantly lower than on 2D CT-image questions. In version B 2D CT-image question scores were also higher than volumetric image test scores. However, the difference was smaller than in version A and not significant. Several discussion points arise from these results. Firstly, it is very difficult to construct different sets of questions with equal difficulty, especially if a set is only composed of 10 or 11 questions, as is the case in the current study. 2D image questions might, unintendedly, be easier. Moreover, different question types were used for 2D and CT-image questions (extended matching versus labeling structures in the image). Thus, differences in score might not be related to the display format alone. Secondly, one 2D CT-image question was very difficult ( $p$ -value .24) compared to the other 2D CT-image questions and might not be representative for 2D CT-image questions in general (outlier). Without this question, the difference in 2D and volumetric CT-image test scores in version B would also have been much larger and comparable with version A scores. Therefore, it seems plausible that the difference in version A 2D and volumetric CT-image scores are more representative. However, this result would benefit from further research. Thirdly, if 2D CT-image questions are indeed less difficult than volumetric image questions an explanation for this finding might be that mastery of more (advanced) knowledge and skills are needed to complete volumetric image questions in comparison to 2D CT-image questions, which is in accordance with the higher reliability of volumetric image question scores. As stated earlier, it is probably easier to find the structure of interest in a 2D image than in a volumetric image, as the structure is already shown. This constitutes a potential risk of underrepresentation of the intended test objectives, which might threaten validity of the test [7].

In contrast to these test score differences we found, participants considered volumetric CT-image questions less difficult than 2D CT-image questions. An explanation of the discrepancy between students' perception and their test scores may be that tracking the course of anatomical structures through volumetric images gives students an unwarranted feeling of confidence. However, the perception of abnormalities in volumetric images with multiple slices might in fact be more difficult than in a single slice, as the abnormality or structure is not directly shown and image manipulating is required [11,12]. Such a false feeling of confidence in volumetric image tests could be compared with the phenomenon of a "false sense of security" reported in, for instance, studies on open-book examinations [13]. In open-book examinations, students have access to their study material, and therefore might experience it as easier than closed-book examinations. However, questions in open-book examinations often assess higher-level cognitive skills, and are in fact more difficult than closed-book exams.

According to the participants, volumetric image questions more accurately reflected the image interpretation process in clinical practice (higher perceived representativeness of the clinical practice) than 2D CT-image questions. This supports the validity of the volumetric approach. However, although all participants were trained in interpreting volumetric images during their radiology

classes, most students in our population had no radiology experience in clinical practice. Further perceived representativeness was evaluated with only one questionnaire item. Repetition of the study with more experienced students (clerks or residents) including a questionnaire with a couple of questions addressing perceived representativeness of volumetric image questions for clinical practice would be useful to evaluate reproducibility of the results to more experienced learners.

Participants of the tests rated the volumetric image testing program as user-friendly. This is an important finding because changing the testing format could increase the complexity of the test, which might impact validity as well as reliability of the test. This threat to validity is called construct irrelevant variance in educational sciences research and implies that test scores can be influenced by other unintended factors (constructs) [7]. For example, tests with a time limit can be extremely difficult and stressful because of the time pressure. The ability of students to cope with stress might then influence test scores as well. The same validity issues can impact the quality of the volumetric test format if the testing program used for volumetric image questions is not intuitive or image manipulation is slow.

Further, an important consideration is that stack viewing of volumetric images are key to cross-sectional modalities, e.g. CT or MRI, and do not play a role in conventional imaging or ultrasound. Depending on the learning objectives, a radiology test should therefore include a mix of 2D images and volumetric images. In the current study we focused on the effect on quality of replacement of 2D CT-images by CT-volumetric images in radiological anatomy tests, and consequently our study results apply only to the testing of CT-interpretation skills.

The current study shows that testing of CT-anatomy knowledge and skills with volumetric images might add to test quality (reliability and perceived representativeness) and that this can be successfully implemented in undergraduate medical education. Though our results are promising, they are still preliminary and might benefit from additional studies. Suggestions for future studies include evaluating the predictive validity of the test showing whether high performers in the volumetric image test indeed interpret radiological images better in clinical practice. Further studies on external validity would be beneficial, to evaluate if scores on the volumetric image test correlate with other tests of radiological knowledge and skills.

## Relevance of article

This article adds to the evidence on how to improve testing of radiological image interpretation skills of students or other trainees in order to stimulate learning and improve radiological performance in clinical practice.

## Role of funding source

The research was partly financially supported by SURF Foundation, Collaborative organisation for ICT in Dutch higher education and research. They had no involvement in the study design, analysis, interpretation of the data or drafting of the article.

## Conflicts of interest

None declared.

## Acknowledgements

The authors wish to thank PJ van der Schoot, ICT manager directorate of Medical Education and Training UMC Utrecht, the

Netherlands, P Simons, evaluation employee of the Center for Research and Development of Education at UMC Utrecht, the Netherlands and E. Zuidema, English teacher at KSG De Breul Zeist, The Netherlands.

#### Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.ejrad.2014.12.015>.

#### References

- [1] Andriole KP, Wolfe JM, Khorasani R, Treves ST, Getty DJ, Jacobson FL, et al. Optimizing analysis, visualization, and navigation of large image data sets: one 5000-section CT scan can ruin your whole day. *Radiology* 2011;259: 346–62.
- [2] Reiner BI, Siegel EL, Siddiqui K. Evolution of the digital revolution: a radiologist perspective. *J Digit Imaging* 2003;16:324–30.
- [3] Krupinski EA, Kundel HL. Update on long-term goals for medical image perception research. *Acad Radiol* 1998;5:629–33.
- [4] Rengier F, Häfner MF, Unterhinninghofen R, Nawrotzki R, Kirsch J, Kauczor HU, et al. Integration of interactive three-dimensional image post-processing software into undergraduate radiology education effectively improves diagnostic skills and visual-spatial ability. *Eur J Radiol* 2013;82:1366–71.
- [5] van der Gijp A, van der Schaaf MF, van der Schaaf IC, Huige JC, Ravesloot CJ, van Schaik JP, et al. Interpretation of radiological images: towards a framework of knowledge and skills. *Adv Health Sci Educ Theory Pract* 2014.
- [6] Messick S. Validity. In: Linn LR, editor. *Educational measurement*. 1989. p. 13–03.
- [7] Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol* 1995;50:741–9.
- [8] Vincken KL, Ravesloot CJ, Mol CP, VQuest. Computer software; Imaging Division University Medical Center Utrecht, The Netherlands, 2010.
- [9] Ebel RL. *Measuring educational achievement*. New Jersey: Prentice-Hall; 1965.
- [10] Pallant J. *Checking the reliability of a scale; SPSS survival manual*. Maidenhead: Open University Press; 2007. p. 95.
- [11] Eckstein M, Thomas J, Palmer J, Shimozaki S. A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. *Percept Psychophys* 2000;62:425–51.
- [12] Palmer J. Attention in visual search: distinguishing four causes of a set-size effect. *Curr Dir Psychol Sci* 1995;4:118–23.
- [13] Dale VH, Wieland B, Pirkelbauer B, Nevel A. Value and benefits of open-book examinations as assessment for deep learning in a post-graduate animal health course. *J Vet Med Educ* 2009;36:403–10.