Cécile J. Ravesloot[a,*], Anouk van der Gijp[a], Marieke F. van der Schaaf, Josephine C.B.M. Huige, Olle ten Cate, Koen L. Vincken, Christian P. Mol and Jan P.J. van Schaik

Identifying error types in visual diagnostic skill assessment

DOI 10.1515/dx-2016-0033
Received September 1, 2016; accepted April 26, 2017

Abstract

Background: Misinterpretation of medical images is an important source of diagnostic error. Errors can occur in different phases of the diagnostic process. Insight in the error types made by learners is crucial for training and giving effective feedback. Most diagnostic skill tests however penalize diagnostic mistakes without an eye for the diagnostic process and the type of error. A radiology test with stepwise reasoning questions was used to distinguish error types in the visual diagnostic process. We evaluated the additional value of a stepwise question-format, in comparison with only diagnostic questions in radiology tests.

Methods: Medical students in a radiology elective (n = 109) took a radiology test including 11–13 cases in stepwise question-format: marking an abnormality, describing the abnormality and giving a diagnosis. Errors were coded by two independent researchers as perception, analysis, diagnosis, or undefined. Erroneous cases were further evaluated for the presence of latent errors or partial knowledge. Inter-rater reliabilities and percentages of cases with latent errors and partial knowledge were calculated.

Results: The stepwise question-format procedure applied to 1351 cases completed by 109 medical students revealed 828 errors. Mean inter-rater reliability of error type coding was Cohen's $\kappa = 0.79$. Six hundred and fifty errors (79%) could be coded as perception, analysis or diagnosis errors. The stepwise question-format revealed latent errors in 9% and partial knowledge in 18% of cases.

Conclusions: A stepwise question-format can reliably distinguish error types in the visual diagnostic process, and reveals latent errors and partial knowledge.

Keywords: assessment; diagnostic errors; image analysis; image interpretation; perception; radiology education; visual diagnosis; visual expertise.

Introduction

Due to increased use of diagnostic imaging [1, 2] and improved accessibility of medical images throughout the hospital, many medical doctors now interpret radiology images on a daily basis. Radiological image interpretation is a complex skill, requiring application and integration of different sorts of knowledge and skills [3], and substantial training and experience to develop [4–6]. Many medical doctors start their careers in the emergency room or on a ward, interpreting radiology images of critically ill patients, while their experience is generally limited. Diagnostic errors can have significant consequences in this acute context [7, 8]. A major part is due to incorrect image interpretation skill of junior doctors [8–11]. Appropriate radiology training and assessment is imperative to improve radiological image interpretation of learners.

Three main components of radiological image interpretation can be distinguished: perception (detection of a lesion), analysis (characterization of a lesion) and synthesis or diagnosis (synthesizing all information into a conclusion or diagnosis) [3]. Specific knowledge and skills are important in each component [3]. For example, accurate perception requires efficient search strategies [12–14], and for correct analysis the ability to characterize findings of abnormalities is essential [3]. Image interpretation errors can occur in all three components and may be caused by different knowledge and skill gaps related to this

[a]Cécile J. Ravesloot and Anouk van der Gijp are joint first authors.
***Corresponding author: Cécile J. Ravesloot,** MD, Radiology Department, University Medical Center Utrecht, E01.132, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands, Phone: +31887556689, Fax: +31302581098, E-mail: C.J.Ravesloot@umcutrecht.nl
Anouk van der Gijp, Josephine C.B.M. Huige and Jan P.J. van Schaik: Radiology Department, University Medical Center Utrecht, Utrecht, The Netherlands
Marieke F. van der Schaaf: Department of Education, Utrecht University, Utrecht, The Netherlands
Olle ten Cate: Center for Research and Development of Education, University Medical Center Utrecht, Utrecht, The Netherlands
Koen L. Vincken and Christian P. Mol: Image Sciences Institute, University Medical Center Utrecht, Utrecht, The Netherlands

component [15–17]. Insight into the nature of errors made by students may help to improve training. Research into diagnostic errors, however, mostly involves eye-tracking studies with a focus on perception errors [12, 18–20]. Little is known about errors beyond perception, like analyzing features of the abnormality and generating a differential diagnosis [3].

Radiology performance tests that are able to trace incorrect image interpretation back to one of the components could be useful to direct feedback, tailor training and stimulate a specific learning process [21]. Radiology tests asking trainees to interpret an image by only asking for a diagnosis might leave perception or analysis errors unattended (latent error). Collateral information or knowledge, such as clinical information or prevalence of the disease, may lead to the correct answer. On the other hand, when the diagnosis is incorrect, perception and analysis might still be correct (partial knowledge).

In this study we evaluated the value of a stepwise reasoning question format in image interpretation assessment. Clinical cases were provided followed by questions regarding perception, analysis and diagnostic skills [3] in a stepwise format. Hypothetically, a set of questions in a stepwise format could be able to unravel the process and provide meaningful performance information such as partial knowledge, latent errors or insights in specific component error types.

Our research questions are (1) Can errors during one of the components of image interpretation (perception, analysis, synthesis) reliably be identified using the stepwise question-format in a radiology performance test? (2) Does the stepwise question approach provide additional performance information compared to single diagnostic questions?

Materials and methods

Study design

We developed a radiology test with stepwise questions, assessing perception, analysis and synthesis skills in image interpretation. The test was administered to 112 medical students in a radiology elective at the University Medical Center Utrecht in the Netherlands from March 2012 to February 2014. The reliability of the stepwise question approach for identifying component errors was estimated with an interrater comparison of scores by two raters. Additional performance information of the stepwise question-format was evaluated by comparing answers on the stepwise questions with the single diagnostic question. The study was approved by the Ethical Review Board of The Netherlands Association of Medical Education (NVMO) and all participants gave informed consent.

Participants

One hundred and twelve medical students took a digital radiology test as a mandatory exam following their radiology clerkship. One hundred and nine of them gave informed consent to analyze their test results and were included in the study. Participants were all fourth to sixth year medical students. All participants had followed an intensive longitudinal radiology course on radiology anatomy and basic radiological knowledge and image interpretation skill on prevalent illnesses as a mandatory part of the undergraduate program. Further, they followed a 6-week radiology elective clinical rotation on knowledge and image interpretation skill of acute and sub-acute diseases in the subareas musculoskeletal, abdominal, chest, and neuroradiology. In addition to the clinical rotation they attended three case-based lessons addressing the radiology subareas.

Instruments

Radiology test: A panel of five radiology education experts created a question bank with image-based cases. Images of acute and subacute diseases were collected and test questions were constructed in the stepwise question-format. The panel also constructed an answer sheet in consensus. A case included one clinical scenario with concomitant imaging and was composed of multiple questions testing perceptual, analytical and diagnostic skills, or knowledge related to the pathology visible on the image (see for an example Figure 1). To assess perception participants were asked to mark the abnormality in the image. Analytical and diagnostic skills were assessed by questions in multiple choice or long menu format. A case did not always include all three components. For some obvious abnormalities, for example a large brain hemorrhage, it was not asked to mark them first. The distribution of the component questions among all completed cases was 40% perception, 22% analysis and 37% diagnosis.

Three versions were composed based on a test matrix: 1A, 2A and 3A. Each test contained 11 to 13 cases in stepwise question-format, including five volumetric CT image cases, five 2D CT image cases and three X-ray image cases, except for test 1A, that contained only one X-ray case in the stepwise question-format. All images were abnormal. Each CT image case was constructed in a volumetric (complete CT-scan with stack viewing) and a 2D format (presenting only selected CT-slices). Of each test, a parallel version was constructed in which all volumetric questions of version A were in 2D format and vice versa: version 1B, 2B and 3B. A scoring model was developed by the expert panel before the first test was administered. The different versions 1A, 2A and 3A were pilot tested by four, one and two medical students, respectively, and if necessary, test questions were rephrased and the scoring model was adjusted.

The tests were administered with VQuest, a digital testing program that allows for radiology image viewing and manipulation, including stack viewing of volumetric CT images in a way that is representative for clinical practice [22, 23]. Participants were allowed to change image contrast and scroll through volumetric images in three viewing directions.

Coding scheme

1. Error types: all errors were coded as related to one of the three components of image interpretation, perception, analysis, and

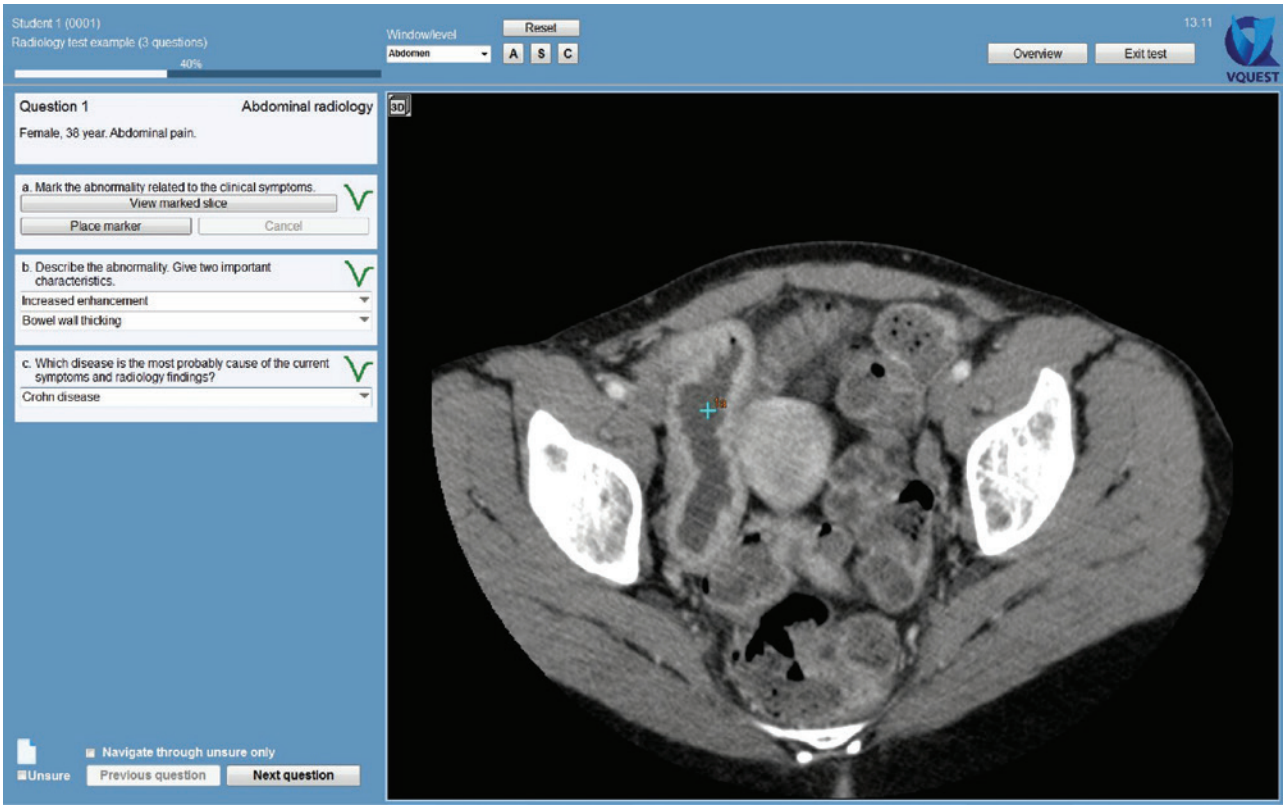

Figure 1: Example of stepwise question-format with questions that aim to test (a) perception, (b) analysis, and (c) diagnosis skill.

synthesis. If an error could not be identified as a component error it was coded as undefined.

2. Additional performance information: at case level we compared the answers on the stepwise question-format questions with the answer on the final diagnostic question (the single diagnostic question, Figure 1, sub question (c) and classified if the stepwise questions provided additional performance information compared to the final diagnostic question alone. Additional performance information was considered to be present when the performance information could only be derived from the answers given on the stepwise questions, and could not be derived from the answers given on the final diagnostic question alone. Each case including at least one error was classified in three categories: (1) Cases with latent errors, being cases in which the student gave the correct diagnosis, while there was either a perception error or analysis error. (2) Cases in which partial knowledge could be valued due to the stepwise question-format. For example, a student marked the correct abnormality, but failed to give the correct diagnosis. (3) Cases in which the stepwise question approach did not reveal latent errors or partial knowledge.

Procedure: The radiology tests were administered in groups of one to seven students. Answers of all participants (on question level) were scored by two researchers (A.G. and C.R., both radiology residents at the time of the study) and discrepancies were solved after reaching consensus. More than one error could be identified within a case. However, if the perception question was incorrectly answered, all other related analytical and diagnostic questions were not marked as separate errors. All identified errors were independently coded and categorized by two researchers (A.G and C.R.). Inter-rater agreement was calculated (with Cohen's κ) per test version after coding all responses. Discrepancies were discussed by the raters until consensus was reached.

Data analysis

Inter-rater agreement for coding errors by two raters was calculated by Cohen's κ per test version to estimate the reliability of error identification.

The percentage of cases with latent errors and partial knowledge were calculated.

Results

Baseline characteristics

In total 16 to 21 students participated per test version which consisted of 11–13 cases. Estimated reliabilities of the tests ranged from low to sufficient on case level (Cronbach's α 0.28–0.69), and from weak to good (Cronbach's α 0.57–0.84) on question level. One hundred and nine

participants completed in total 1351 cases. In 715 completed cases no errors were identified. In the remaining 636 cases 828 errors were found. The average number of errors per case was 0.6 (SD 0.7). The median test score on the stepwise questions was 77.8% with an interquartile range of 15.4%.

Inter rater agreement

Cohen's κ for classification of error types and additional performance information are given in Table 1.

Error types

In total 651 errors could be identified as a component error (79%). The remaining 21.4% was left undefined. The distribution of the component errors was 38%–20%–41% and is comparable with the distribution of the questions per component in the test.

Additional performance information

In 27.2% of the cases (367 of 1351 cases) the stepwise question-format revealed latent errors or partial knowledge, that could not be derived from the final diagnostic question alone. In 18% of the cases (243 cases) participants were evaluated as being a more competent interpreter based on the stepwise approach, because the stepwise questions revealed partial knowledge, even though they provided a wrong diagnosis. For example, a participant gave an incorrect diagnosis (a Jones fracture instead of a tear drop fracture), but correctly indicated and analyzed the lesion, as shown by the answers on the preceding questions (see Figure 2B). In 9.3% of the cases (125 cases) the diagnosis was correct, but the stepwise questions showed an error in either perception (60 cases) or analysis (65 cases) of the abnormality, signifying a latent error. For example, a participant gave the correct diagnosis Crohn disease, but failed to correctly mark the abnormal bowel loop (see Figure 2A). Another participant correctly diagnosed a patient with pneumonia, however did not mark the pneumonia, but indicated normal lung tissue as abnormal. The case text "dyspnea and fever" and the participant's knowledge about symptoms of pneumonia possibly resulted in the correct answer. These participants' image interpretation performances were valued as being lower, due to the latent errors revealed by the stepwise question-format. As shown in Figure 3 in 19.8% of the cases (268 cases) the stepwise question-format revealed no partial knowledge or latent errors.

Table 1: Inter-rater agreement for coding error types and additional performance information by two raters, estimated by Cohen's κ per test version.

	Error types		Additional performance information	
	Version A	Version B	Version A	Version B
Test 1	0.73	0.86	0.66	0.81
Test 2	0.77	0.88	0.84	0.95
Test 3	0.68	0.90	0.63	0.81

Discussion

The stepwise question-format reliably distinguished errors related to the image interpretation process of medical students in a radiology elective. Almost 80% of the errors could be related to an error in one of the components of image interpretation.

The percentages of error types should be interpreted with caution and should not be interpreted as actual prevalence of errors, because they are partly test-driven due to variation in the stepwise questions. The distribution of the different component errors (39%–20%–41%) was similar to the distribution of the stepwise questions on perception, analysis, and diagnosis (40%–22%–38%), so all errors seemed to occur almost with equal frequency in proportion to the questions that provoke the errors. However, this result cannot be compared with the relatively high proportion of perceptual errors in clinical practice reported in the literature [15, 17], because the current test only included abnormal cases. Therefore students were biased in the sense that they knew that an abnormality was present, which differs from clinical practice in which many cases are normal.

In more than a quarter of all cases the stepwise question-format revealed additional performance information of the learner that could not be derived from the final diagnostic question alone. In almost two-thirds of these cases partial knowledge was uncovered, while in one-third of the cases flaws in perception or analysis appeared which would have stayed unnoticed if only a diagnosis was asked.

These results indicate that the stepwise question-format provides the teacher with additional performance information compared to only diagnostic questions

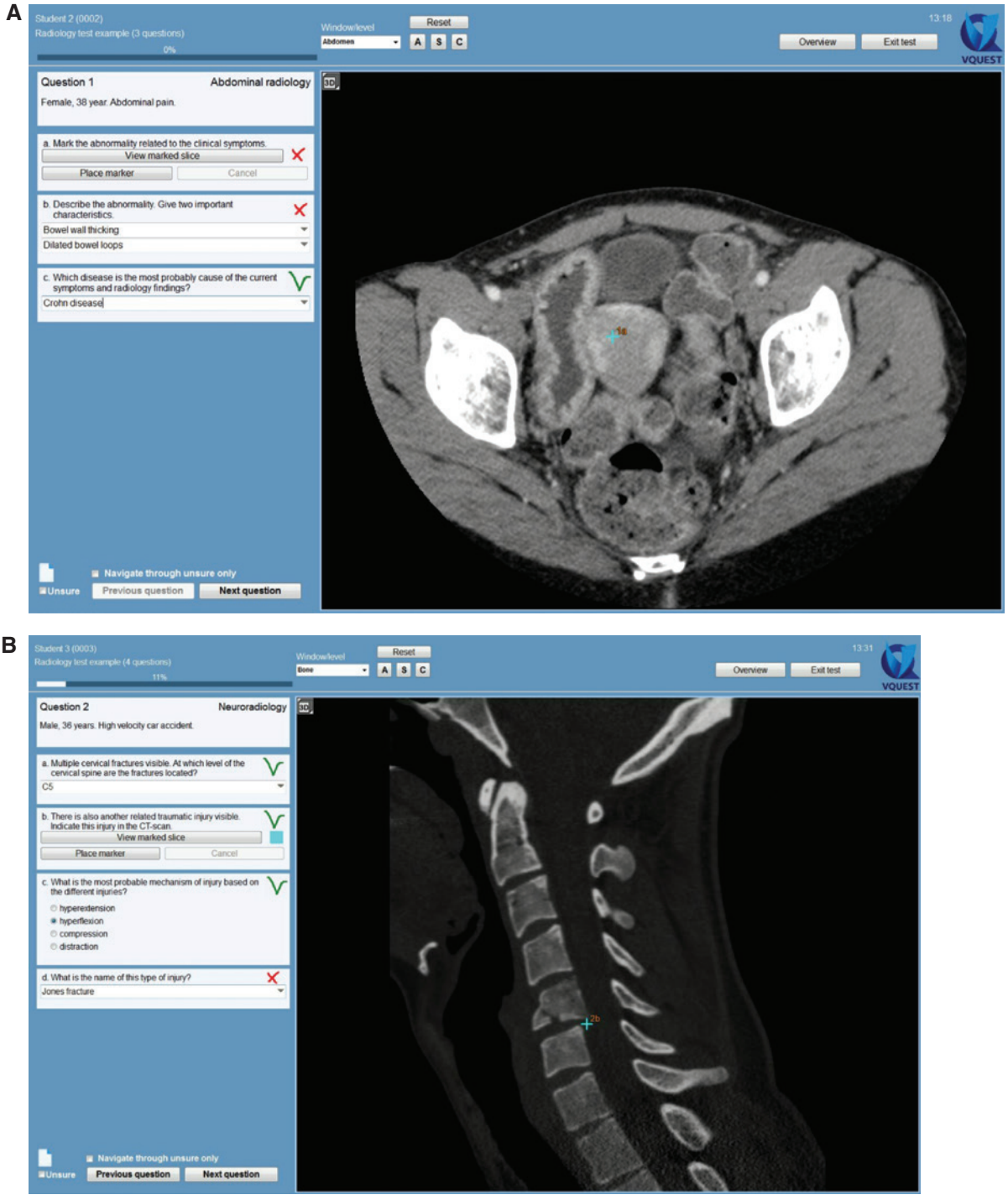

Figure 2: Examples of additional performance information of the stepwise question-format by showing latent errors (A) or partial knowledge (B).
(A) Latent error. This student gave the correct diagnosis "Crohn disease", but marked a normal structure (uterus) as the abnormality (perception error). (B) Partial knowledge. This student gave the incorrect diagnosis: "Jones fracture" instead of "teardrop fracture" (diagnosis error), however the answers on the other questions, addressing perception and analysis, revealed that the student interpreted the image otherwise correctly.

formats, because this method can reliably identify in which component of the image interpretation process the student succeeded or failed. Stepwise questions give a more detailed and probably more accurate representation of the performance of students and uncovers latent errors in the image interpretation process.

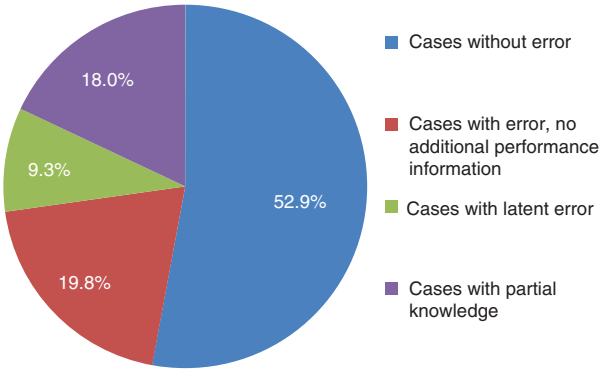

Figure 3: Percentage of cases with additional performance information derived from the stepwise question-format.

The additional information can be useful for both teachers and students. Teachers will get more insight into the errors of students and their actual performance level which may be used to inform individual feedback to learners. Students may use the information to detect their own strengths and weaknesses, and to monitor their development in specific image interpretation skills.

Some limitations of the study should be addressed. Image interpretation is a complex process and the distinction between error types is rather artificial. Perception, analysis and synthesis processes are highly interrelated and constantly alternate within the image interpretation process [24]. The stepwise question approach directed the diagnostic reasoning process into a stepwise reasoning process, broken down in three components, and therefore allowed for distinguishing three components of the image interpretation process. Breaking down errors in these three components should only be recognized as one of many ways of specifying errors in diagnostic radiology for an improved understanding of decision-making.

A drawback of the stepwise question method is the increased time needed to complete the tests. Single questions would allow for a larger number of cases in the same testing time, which could increase reliability. The reliability of the current test ranged from low to sufficient on case level, probably because of the low number of cases [11–13] in each test version. It is important to take the purpose of the test into account, when deciding between either the stepwise question-format or only diagnostic questions. The information revealed by the stepwise question-format approach can be valuable for feedback in formative testing, while a large number of questions can be necessary to increase reliability in summative tests.

In this study, we only applied the stepwise question method to medical students and we cannot generalize our results to other levels of expertise. For example, the error rate and distribution of error types will likely vary with knowledge and experience levels. Testing the method in residents or radiologists is subject to further research.

We conclude that the stepwise question approach succeeded in identifying specific error types in the image interpretation process. Besides, it revealed partial knowledge and latent errors. Both can be used to tailor image interpretation training. Specifying error types may be advantageous for giving effective feedback, because a specific task evaluation is one of the characteristics of high-quality feedback [25, 26].

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: SURF Foundation (Grant Number: TTL 11.0269).
Employment or leadership: None declared.
Honorarium: None declared.
Competing interests: The funding organization(s) played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

References

1. Bhargavan M, Sunshine JH. Workload of radiologists in the United States in 2002–2003 and trends since 1991–1992. Radiology 2005;236:920–31.
2. Bhargavan M, Kaye AH, Forman HP, Sunshine JH. Workload of radiologists in United States in 2006–2007 and trends since 1991–1992. Radiology 2009;252:458–67.
3. 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;19:565–80.
4. Nodine CF, Kundel HL, Mello-Thoms C, Weinstein SP, Orel SG, Sullivan DC, et al. How experience and training influence mammography expertise. Acad Radiol 1999;6:575–85.
5. Norman GR, Coblentz CL, Brooks LR, Babcook CJ. Expertise in visual diagnosis: a review of the literature. Acad Med 1992;67:S78–83.
6. Taylor PM. A review of research into the development of radiologic expertise: implications for computer-based training. Acad Radiol 2007;14:1252–63.
7. Gruen RL, Jurkovich GJ, McIntyre LK, Foy HM, Maier RV. Patterns of errors contributing to trauma mortality: lessons learned from 2594 deaths. Ann Surg 2006;244:371–80.

8. Wechsler RJ, Spettell CM, Kurtz AB, Lev-Toaff AS, Halpern EJ, Nazarian LN, et al. Effects of training and experience in interpretation of emergency body CT scans. Radiology 1996;199:717–20.
9. Guly HR. Diagnostic errors in an accident and emergency department. Emerg Med J 2001;18:263–9.
10. Rhea JT, Potsaid MS, DeLuca SA. Errors of interpretation as elicited by a quality audit of an emergency radiology facility. Radiology 1979;132:277–80.
11. Gwynne A, Barber P, Tavener F. A review of 105 negligence claims against accident and emergency departments. J Accid Emerg Med 1997;14:243–5.
12. Hu CH, Kundel HL, Nodine CF, Krupinski EA, Toto LC. Searching for bone fractures: a comparison with pulmonary nodule search. Acad Radiol 1994;1:25–32.
13. Krupinski EA. Visual scanning patterns of radiologists searching mammograms. Acad Radiol 1996;3:137–44.
14. Drew T, Le-Hoa Vo M, Olwal A, Jacobson F, Seltzer SE, Wolfe JM. Scanners and drillers: characterizing expert visual search through volumetric images. J Vis 2013;13:pii: 3.
15. Renfrew DL, Franken Jr EA, Berbaum KS, Weigelt FH, Abu-Yousef MM. Error in radiology: classification and lessons in 182 cases presented at a problem case conference. Radiology 1992;183:145–50.
16. Pinto A, Acampora C, Pinto F, Kourdioukova E, Romano L, Verstraete K. Learning from diagnostic errors: a good way to improve education in radiology. Eur J Radiol 2011;78:372–6.
17. Donald JJ, Barnard SA. Common patterns in 558 diagnostic radiology errors. J Med Imaging Radiat Oncol 2012;56:173–8.
18. Donovan T, Litchfield D. Looking for cancer: expertise related differences in searching and decision making. Appl Cognit Psychol 2013;27:43–9.
19. Rubin GD, Roos JE, Tall M, Harrawood B, Bag S, Ly DL, et al. Characterizing search, recognition, and decision in the detection of lung nodules on CT scans: elucidation with eye tracking. Radiology 2015;274:276–86.
20. Kundel HL, Nodine CF, Carmody D. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. Investigat Radiol 1978;13:175–81.
21. Pecaric M, Boutis K, Beckstead J, Pusic M. A big data and learning analytics approach to process-level feedback in cognitive simulations. Acad Med 2017;92:175–84.
22. Ravesloot CJ, van der Schaaf MF, van Schaik JP, ten Cate OT, van der Gijp A, Mol CP, et al. Volumetric CT-images improve testing of radiological image interpretation skills. Eur J Radiol 2015;84:856–61.
23. Ravesloot CJ, van der Gijp A, van der Schaaf MF, Huige JC, Vincken KL, Mol CP, et al. Support for external validity of radiological anatomy tests using volumetric images. Acad Radiol 2015;22:640–5.
24. Morita J, Miwa K, Kitasaka T, Mori K, Suenaga Y, Iwano S, et al. Interactions of perceptual and conceptual processing: expertise in medical image diagnosis. Int J Hum Comp Stud 2008;66:370–90.
25. Sadler DR. Beyond feedback: developing student capability in complex appraisal. Assess Eval High Educ 2010;35:535–50.
26. Black P, William D. Assessment and classroom learning. Assess Educ Princ Pol Pract 1998;5:7–74.