Increasing Authenticity of Simulation-Based Assessment in Diagnostic Radiology

Anouk van der Gijp, MD, PhD;
Cécile J. Ravesloot, MD, PhD;
Corinne A. Tipker, MSc;
Kim de Crom, MSc;
Dik R. Rutgers, MD, PhD;
Marieke F. van der Schaaf, PhD;
Irene C. van der Schaaf, MD, PhD;
Christian P. Mol, MSc;
Koen L. Vincken, PhD;
Olle Th.J. ten Cate, PhD;
Mario Maas, MD, PhD;
Jan P.J. van Schaik, MD, PhD

Introduction: Clinical reasoning in diagnostic imaging professions is a complex skill that requires processing of visual information and image manipulation skills. We developed a digital simulation-based test method to increase authenticity of image interpretation skill assessment.
Methods: A digital application, allowing volumetric image viewing and manipulation, was used for three test administrations of the national Dutch Radiology Progress Test for residents. This study describes the development and implementation process in three phases. To assess authenticity of the digital tests, perceived image quality and correspondence to clinical practice were evaluated and compared with previous paper-based tests (PTs). Quantitative and qualitative evaluation results were used to improve subsequent tests.
Results: Authenticity of the first digital test was not rated higher than the PTs. Test characteristics and environmental conditions, such as image manipulation options and ambient lighting, were optimized based on participants' comments. After adjustments in the third digital test, participants favored the image quality and clinical correspondence of the digital image questions over paper-based image questions.
Conclusions: Digital simulations can increase authenticity of diagnostic radiology assessments compared with paper-based testing. However, authenticity does not necessarily increase with higher fidelity. It can be challenging to simulate the image interpretation task of clinical practice in a large-scale assessment setting, because of technological limitations. Optimizing image manipulation options, the level of ambient light, time limits, and question types can help improve authenticity of simulation-based radiology assessments.
(*Sim Healthcare* 12:377–384, 2017)

Key Words: Simulation-based assessments, clinical reasoning, medical images, radiology.

Complex clinical skills are challenging to assess. The closest reflection of actual clinical behavior is obtained by workplace-based assessments in which trainees are judged for their performance at clinical tasks, without standardization of patients or settings.[1] However, the inherent lack of standardization and the difficulty to attain high levels of interrater agreement can diminish the reliability of workplace-based assessments.[2] In addition, workplace-based assessments generally require direct observation and much time and effort of faculty members. Simulation-based assessment aims to test participants' clinical performance in a standardized setting. Most medical simulations use virtual patients,[3,4] such as mannequins, standardized patients, or laparoscopy simulators. These patient models attempt to approach the experience with actual patients. In visual diagnostic domains, such as radiology, the imaging data of actual patients can be used to simulate the clinical task of image interpretation.

Although the use of actual patient data adds to fidelity of the simulation, it does not guarantee a high level of authenticity.[5] Authenticity refers to the degree to which an assessment resembles the task in professional practice.[6] To simulate a clinical task in a credible way, a certain level of authenticity is needed. Much of current radiological practice involves reading volumetric images. Volumetric images are sets of successive cross-sections of a human body part. These cross-sections are usually magnetic resonance (MR) or computed tomography (CT) images. Radiologists scroll through the set of images to detect lesions and diagnose diseases. Advanced image interaction or manipulation tools can be used, such as scrolling through images in any direction and adjusting contrast settings.[7–9] Interpreting volumetric images requires different cognitive skills than interpreting two-dimensional (2D) images[10] and requires the processing of large amounts of visual data. These skills cannot be adequately captured in a paper-based test (PT). Therefore, a computer-based test is needed to include human-computer interactions and to reach an acceptable level of authenticity. However, a higher resemblance of clinical practice does not

From the Department of Radiology (A.V.D.G., C.J.R., D.R.R., I.C.V.D.S., C.P.M., K.L.V., J.P.J.V.S.), University Medical Center, Utrecht; Department of Radiology (C.A.T., K.D.C., M.M.), Academic Medical Center, Amsterdam; Examination Committee (D.R.R.), Radiological Society of the Netherlands; Department of Education (M.F.V.D.S.), Utrecht University; and Center for Research and Development of Education (O.T.J.T.C), University Medical Center, Utrecht, the Netherlands.

Reprints: Anouk van der Gijp, MD, PhD, Radiology Department UMC Utrecht, E01.132, Heidelberglaan 100, 3584 CX Utrecht, the Netherlands (e-mail: A.vanderGijp-2@umcutrecht.nl).

Supported by SURF, the higher education and research partnership organization for Information and Communications Technology. For more information about SURF, please visit www.surf.nl.

The authors declare no conflict of interest.

DOI: 10.1097/SIH.0000000000000278

necessarily mean that examinees experience this as such, because authenticity of a test is partly in the eyes of the beholder.[11]

To evoke true radiological diagnostic reasoning, not only the viewing mode but also the cognitive characteristics of the test should align with clinical practice. In clinical reasoning literature, many question types have been developed and investigated in an attempt to test clinical reasoning.[12] The diagnostic process in visual domains primarily focuses on hypothesis generation, based on image characteristics and available clinical information. Questions should contain images, accompanied with limited clinical information. A response format that requires the active generation of a diagnosis aligns best with the clinical task of hypothesis generation. The perceptual component may be captured with question types that test detection skills, such as asking for marking abnormalities or anatomical structures.

Many initiatives to simulate the radiological image interpretation task have been reported in radiology education literature. Most simulations are used for e-learning purposes.[13,14] The image interaction possibilities in these simulations are usually absent or limited,[14] and learners tend to provide feedback suggesting increased interactions with radiographic images.[15–17] Some studies suggest that introducing teaching material with image interaction possibilities has a positive effect on learning outcomes,[18,19] but the level of evidence is low. The use of simulations for radiology assessments is less widely reported, and image interaction possibilities are lacking or limited.[20,21]

In a previous study, we found that the introduction of volumetric images has the potential to improve the validity of radiology anatomy tests in medical students.[22] According to the medical students, testing with volumetric images reflected clinical practice better than testing with 2D images.[22] This population had no experience in clinical practice and the questions only involved normal anatomy, whereas clinical radiology involves recognition and interpretation of pathological images. These results should therefore be verified in a population that has a good understanding of radiology practice and with questions that aim to test image interpretation in pathological cases as well.

The purpose of this study was threefold:

1. To develop and implement a digital simulation-based assessment method for monitoring image interpretation skills of radiology trainees.
2. To describe the methods, challenges, and lessons learned from this development and implementation process.
3. To evaluate the authenticity of the digital test in comparison with former paper-based assessments.

METHODS

Setting

In the Netherlands, radiology residency involves 5 years of training in academic and nonacademic hospitals. The Dutch Radiology Progress Test (DRPT) is a semiannual mandatory test for residents. The DRPT aims to test development of radiological knowledge: all residents, regardless of their level of experience, take the same end-of-training level test. The DRPT has a formative purpose, which is to provide feedback to trainees, to reflect their progress and guide self-directed learning, rather than to yield a summative score.[23] The questions are constructed by expert radiologists of the examination committee of the Radiological Society of the Netherlands. The test has been described in more detail in a previous study.[24] After 10 years of paper-based testing, the DRPT was transformed into a digital format in 2013. In April 2013, a pilot test was conducted among 383 participants. The aim of this pilot was to examine the feasibility of the test procedure and its technical performance. The development and implementation process of the following three digital radiology progress tests [signified below as digital test (DT) 1, 2, and 3], administered in 2013 and 2014, is described and evaluated in the current study. To compare the DT with previous paper-based testing, the three most recent PTs before the transition (PT 1, 2, and 3, administered in 2011 and 2012) were used. The current study focuses on the image-based questions assessing image interpretation skills.

Study Design

Three-Phase Developmental Process Evaluation

We longitudinally describe the development and implementation process of the three DTs. In the method section of DT 1, the initial implementation of the DT method will be outlined. The method section of DT 2 and DT 3 will focus on the changes that were made based on the results of the previous DTs. Both quantitative and qualitative results informed decisions for further improvement of the subsequent DTs. In the development process, we focused on improving authenticity of the image interpretation task. Radiology expertise literature distinguishes visual and cognitive components of image interpretation.[25–28] Therefore, we distinguish viewing task and cognitive task characteristics to evaluate task authenticity. Viewing task characteristics include aspects of the task that are related to the images, that is how images are displayed and to what extent they can be manipulated. Cognitive task characteristics involve aspects of the task that are related to the thinking process of the trainee, which is processing the visual information, and diagnostic reasoning. After each DT, we evaluated the participants' perceptions about the authenticity of the digital image questions compared with former paper-based questions. In addition, we compared the reliability of the digital image question subtests with former paper-based versions.

Participants

The three digital progress tests were taken by 356 (DT 1), 367 (DT 2) and 349 (DT 3) Dutch radiology residents. The mean duration of radiology training of the participants at the time of testing was 2.4, 2.5, and 2.4 years. The PTs were taken by 357 (PT 1), 367 (PT 2), and 354 (PT 3) residents, who had completed an average duration of radiology training of 2.3 years across all three tests.

Questionnaire

Within 1 week after the DTs, participants received an invitation to answer an online questionnaire with a reminder after 1 week. Questions concerned the perceived correspondence to clinical practice and image quality of the digital and paper-based image questions. Questions concerning the paper-based image questions were only completed by respondents who had previously completed a paper-based version of the test. Response format was on a five-point Likert scale, ranging from

"1 = insufficient" to "5 = good." Participants were asked to provide suggestions for improvement in four open comment sections of each questionnaire. Open comments regarding the image questions were used for qualitative analysis to explain or complement the quantitative findings. The questionnaire also included questions about the logistics of the test administration, the user-friendliness of the test application, and the test environment, the answers of which were to be used to guide the improvement of subsequent tests.

Response rates were 52%, 46%, and 43% after DT 1, 2, and 3, respectively. The distribution of participants over the training years across all three questionnaires together was the following: year 1, 20%-24%; year 2, 20%-21%; year 3, 15%-25%; year 4, 20%-27%; and year 5, 10%-16%.

Statistical Analysis

For comparing the reliability of the digital image questions with paper-based image questions, Cronbach α was calculated for each subtest level, after removal of flawed questions (3%-5%) determined so by the examination committee and guided by item analysis. To compare reliabilities of the different subsets of image questions, Spearman Brown formula was applied to correct for test length differences.[29] Item-total correlations (r_{it}) of the image questions were calculated.

After assumption checks, paired *t* tests and Wilcoxon signed-rank tests were conducted to compare survey ratings concerning digital testing with those concerning paper-based testing. For qualitative analysis for the purpose of this study, comments concerning the image-based questions were categorized by perceptions regarding viewing task characteristics and regarding cognitive task characteristics. Themes concerning these characteristics were identified and reported when three or more comments were related to a particular theme. Comments given twice or more by the same participant were counted only once. Comments concerning the nonimage-based questions and the organization of the test were analyzed separately and used for improvement of subsequent tests.

Ethical Approval

The ethical review board of the Netherlands Association for Medical Education approved the study (ERB number 206).

Test Format

The first DT contained 200 questions, equal to the PT versions. The number of image-based questions was 36 and comparable with previous PTs. In the first DT, 56% of the image-based questions contained volumetric images. In the scoring model of the DTs, each correct answer to a question yielded one point. The scoring model of the PTs was based on formula scoring, because the questions included a "don't know" answer option. In formula scoring, scores are calculated by subtracting the number of wrong answers from the number of correct answers to correct for guessing. The "don't know" answer option was removed before the DT implementation, based on a previous experiment about the effect of the "don't know" answer option.[30]

Test Environment

A DT environment, VQuest (http://www.vquest.eu or http://vquest.bluefountain.nl/en/), was used to administer the digital DRPT. The test application was developed at University Medical Center Utrecht and was used in previous studies to improve radiology tests for medical students.[22,31] It allows for volume data set viewing and image manipulation. Participants can navigate through volumetric images in different viewing directions. The program also allows for zooming in and out and adjusting image contrast. Examples of a volumetric image question are shown in Figures 1 and 2. A video of the

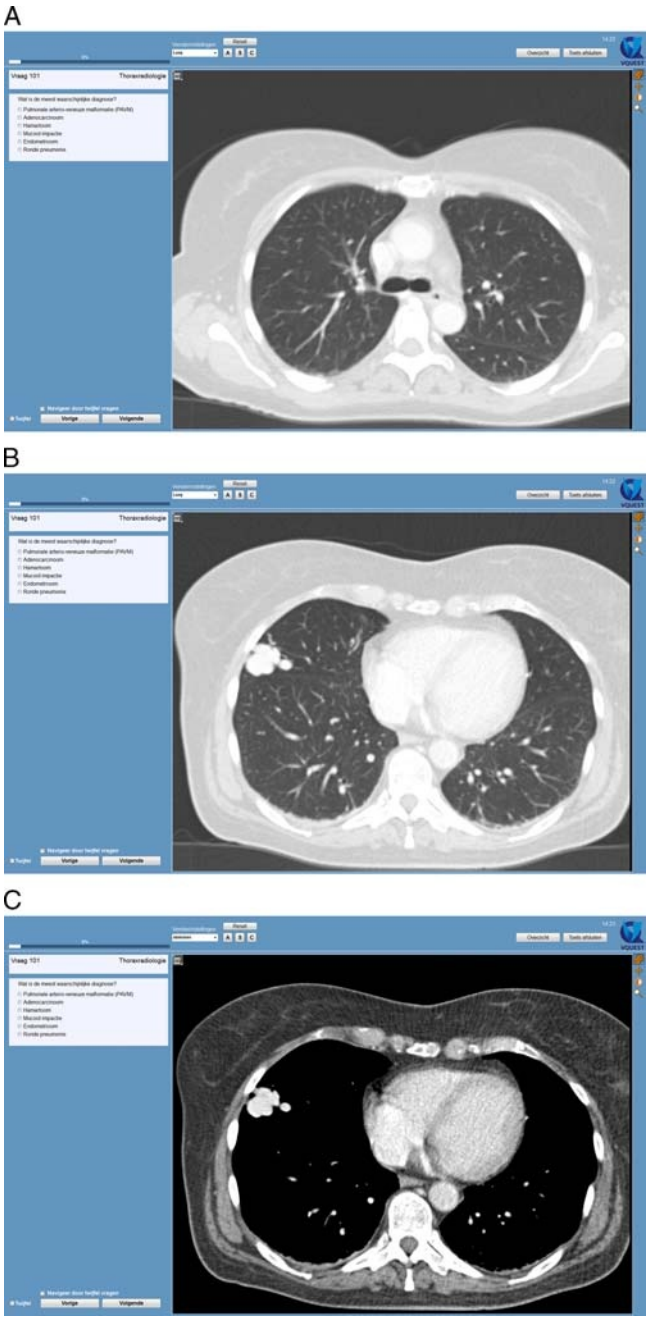

FIGURE 1. A, An example of a multiple choice volumetric image question. This is a chest CT scan with an abnormality. Participants are asked to choose the correct diagnosis. The abnormality is not visible on this cross-section, so participants have to scroll through the set of images to detect and analyze the abnormality. B, The same question as in Figure 1A, showing a different cross-section of the CT scan. On this cross-section, the abnormality is visible. C, The same question as in Figures 1A and B, showing the same cross-section as in Figure 1B, but with a different contrast setting. Changing contrast can facilitate image interpretation, for example, by showing the tissue characteristics of the abnormality.

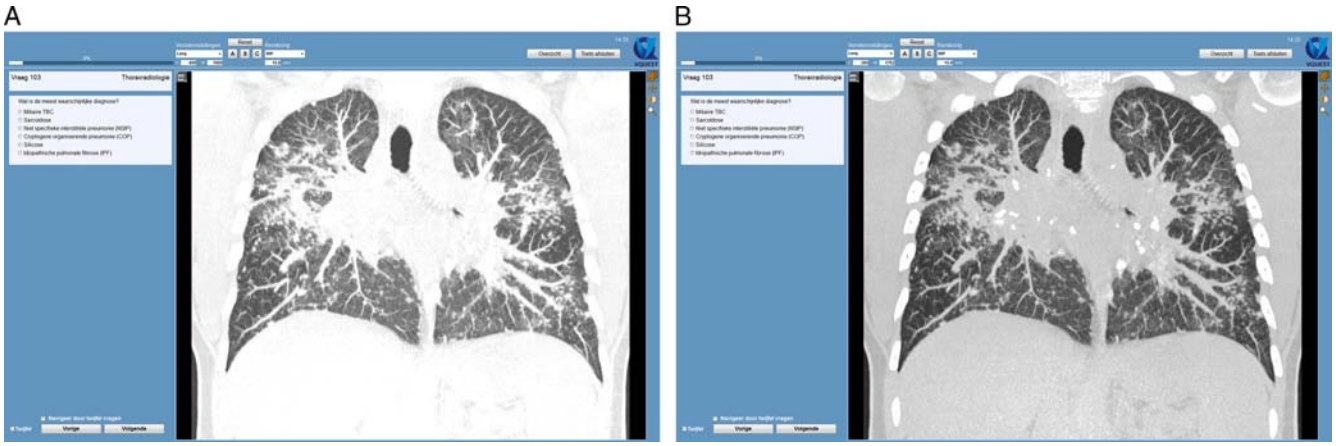

FIGURE 2. A, Another example of a multiple choice volumetric image question. This cross-section of a chest CT scan is shown in a different viewing direction than the cross-section in Figure 1. In addition, an advanced image reconstruction is applied: a maximum intensity projection. This reconstruction method can be useful for detection of lung nodules or for showing patterns of lung nodules. B, The same question as Figure 2A, but with different contrast settings. This example illustrates that some structures become only visible after changing contrast settings. For example, the ribs become visible and multiple calcifications can be observed (bright white structures in the center of the image).

test application is available in the supplementary of previously published work.[22]

Simulating Viewing Task Characteristics

To simulate the viewing characteristics of the task, we listed the viewing characteristics of the task in clinical practice and the PT. Within the possibilities and restrictions of the available hardware and software, we implemented viewing characteristics in the DT that were closest to clinical practice. Because of screen size and resolution restraints, the maximum number of images displayed at once was restricted. Even though ample image manipulation options were available in the test environment, not all options were included to their full extent. We anticipated that too many options could overwhelm participants and could increase reading time. Many manipulation options were therefore only made available if considered to have added value in a particular question. The viewing task characteristics of the first DT, the PT, and clinical practice are described in Table 1.

Simulating Cognitive Task Characteristics

To improve simulation of the diagnostic reasoning task, hotspot and multiple choice questions were introduced, in addition to the true/false questions. The hotspot question aimed to test perception skills,[9] by asking participants to place a marker in an abnormality or an anatomical structure. The multiple choice question could be used to test analysis or synthesis skills,[9] for example, testing the ability to diagnose, by listing a number of possible diseases.

RESULTS

Phase 1: Evaluation of DT 1

Item Analysis

Reliability of the digital image question subtest of DT 1 was comparable with the reliability of the paper-based image questions when corrected to a 60 image-based questions test with Spearman Brown formula (Table 2). Average item-total correlation (r_{it}) values for digital image questions per question type are given in Table 3.

Viewing Task Characteristics

Image quality was rated significantly lower than the former paper-based image quality [mean(SD) = 2.7 (1.2, n = 143) and mean (SD) = 3.1 (1.1, n = 143), respectively, $t(142) = -2.84$, $P < 0.05$, $d = 0.35$]. Qualitative data analysis showed five themes with respect to the suggestions for improvement of the viewing task. The most prevalent theme was "image manipulation options," with most comments being related to the desire to zoom in and out. The second most important theme was "image resolution/size" due to complaints about the small screen size

TABLE 1. Viewing Task Characteristics in Clinical Practice, the Digital Image Questions of DT 1, and the Paper-Based Image Questions

Viewing Task Characteristics	Clinical Practice	DT 1	PTs
Image display			
Screen size	Typically 20–30 in	15.6 in	NA
Screen resolution	Typically 1600 × 1200–3280 × 2048	1366 × 768	NA
Display >1 image at once	Yes, multiple	Yes, maximum of 4 images	Yes, multiple
Comparing with previous images	On demand	Only when provided in the question	Only when provided in the question
Image manipulation			
Scrolling back and forth	Yes	Yes	No
Changing contrast setting	Yes, presets or free in any direction	Yes, presets or free in any direction	No
Changing viewing direction	Yes, presets or free in any direction	Presets, when considered to have added value	No
Zooming in and out	Yes	No	No
Making advanced reconstructions	Yes	Yes, when considered to have added value	No

NA, not applicable.

TABLE 2. Reliabilities of Digital and Paper-Based Image Questions

	PT 1	PT 2	PT 3	DT 1	DT 2	DT 3
k (image items)	37	37	36	36	40	60
Reliability (Cronbach α)	0.67	0.78	0.76	0.74	0.72	0.80
Spearman Brown corrected α ($k = 60$)	0.77	0.85	0.84	0.83	0.79	0.80

and image resolution. In addition, the ambient lighting was criticized for being too high, leading to hindering reflections on the computer screens. The number of comments for each theme is listed in Table 4.

Cognitive Task Characteristics

Seventy-four percent of the participants agreed that digital radiology progress testing corresponds better to clinical practice than paper-based testing. The participants' perceptions of correspondence with clinical practice between the image questions in DT 1 and the former paper-based image questions did not differ. Qualitative data analysis revealed that reflections upon the cognitive task characteristics were very scarce, only seven comments, and the predominant view was that the digital 2D and volumetric image questions reflected clinical practice better than paper-based image questions. For example, one of the participants wrote: *"Image questions with full data sets fit the reality of practice better."* Next, the participant added: *"However, the quality of the images especially the x-rays, should be taken care of."*

Phase 2: Improvement and Evaluation of DT 2

Test Improvements

Improving actual image quality was not possible because of hardware limitations. In an attempt to improve the perceived image quality, several adjustments were made in the viewing characteristics of the second test: (1) the zooming option was made available, (2) a full screen option was implemented that enabled double clicking on of the images for a full screen view of the image, and (3) ambient light was reduced.

To compensate for the increase in time investment involved in answering volumetric image questions, the total number of questions was decreased from 200 to 180. Besides, the proportion of volumetric image questions was reduced to 40% of 40 image-based questions.

Item Analysis

Reliability of the digital image question subtest of DT 2 was comparable with the reliability of the paper-based image questions when corrected to a 60 image-based questions test with Spearman Brown formula (Table 2). Average item-total correlation (r_{it}) values for digital image questions per question type are given in Table 3.

Viewing Task Characteristics

Despite the attempts to improve perceived image quality, digital image quality was still rated significantly lower than the former paper-based image quality [mean(SD)= 2.7 (1.1, n = 129) and mean (SD) = 3.1 (1.1, n = 129), respectively, $t(128) = -2.99$, $P < 0.05$, $d = 0.36$]. In the qualitative data analysis, the emphasis was again on the image manipulation options and the image size, although also scrolling speed was subject of discussion (Table 4). Although the number of comments had decreased, there was still criticism on the screen size and image resolution. This was specifically related to the questions with mammography images, x-rays of the breasts, which can contain very small calcifications. One of the comments: *"The details on the x-ray images, such as micro calcifications, are very hard to distinguish on these computer screens."* The comments on the image manipulation options had decreased and changed direction. The zooming option was now available, but it was not always functioning well. The number of comments about ambient lighting had decreased from 34 to 6. In addition, comments about the scrolling speed had increased; slow and faltering scrolling was experienced.

TABLE 3. Average r_{it} Values of Digital Image Questions per Question Type

r_{it} Values Image Questions	TFQ (*k*)	MCQ (*k*)	HSQ (*k*)	LMQ (*k*)
DT 1	0.15 (8)	0.26 (24)	0.31 (4)	—
DT 2	0.18 (12)	0.32 (24)	0.34 (4)	—
DT 3	0.22 (16)	0.30 (32)	0.27 (4)	0.35 (8)

HSQ, hotspot question; *k*, number of questions; LMQ, long-menu question; MCQ, multiple choice question; TFQ, true/false question.

TABLE 4. Number of Comments Related to the Viewing Task, Categorized in Five Themes

Theme	DT 1	DT 2	DT 3
No. responders	186	169	152
Scrolling speed	4	32	30
Loading speed of images	3	4	19
Image resolution/size	78	31	11
Image manipulation options	112	37	10
Ambient light reflection on screen	34	6	1

Cognitive Task Characteristics

Seventy-two percent of the participants agreed that digital radiology progress testing corresponds better to clinical practice than paper-based testing. According to the participants, digital image questions of the second test corresponded significantly better to clinical practice than paper-based image questions [mean (SD) = 3.3 (1.0, n = 132) and mean (SD) = 3.1 (1.0, n = 132), respectively, $t(131) = 1.99$, $P < 0.05$, $d = 0.20$].

Again, the qualitative data reflected the dominant view that digital 2D and volumetric image questions reflect clinical practice better than paper-based image questions. In addition, some participants shared the opinion that the value of testing with volumetric images is a trade-off between improved reflection of clinical practice and increased time needed to complete the questions. For example, one of their comments was: *"Changes in viewing direction and contrast settings enable better interpretation of the abnormalities on the image, but it takes a lot of time."*

Phase 3: Improvement and Evaluation of DT 3

Test Improvements

To improve the viewing task characteristics, the software was optimized to improve the speed of image manipulation, especially the zooming option and the scroll function. The examination committee received extra instructions about what type of images and cases were preferable or not. For example, it was recommended not to include too large volumetric data sets and only include the relevant parts of the image, because too large data sets could negatively affect the performance of the software, such as loading time and scrolling speed.

To further enhance the test goal of image interpretation skills, the number of image-based questions was increased to

60 (of which 38% volumetric images). Besides, the options of question types were expanded with a long-menu question. This question type requires selecting an answer from a long list of answer options, which only appears after typing two or more letters corresponding to the available options. The long-menu question was introduced to improve authenticity of the simulation, because there are no multiple choice options in clinical practice, and diagnoses have to be actively generated.

Item Analysis

Reliability of the digital image question subtest of DT 3 was comparable with the reliability of the paper-based image questions when corrected to a 60 image-based questions test with Spearman Brown formula (Table 2). Average item-total correlation (r_{it}) values for digital image questions per question type are given in Table 3.

Viewing Task Characteristics

After these improvements, image quality of the third DT was rated significantly higher than the former paper-based image quality [mean (SD) = 3.1 (1.1, n = 109) and mean (SD)= 2.7 (1.1, n = 109), respectively, $t(108) = 2.91$, $P < 0.01$, $d = 0.36$]. The number of comments about the image manipulation options and image size had decreased significantly, and there was only one complaint left about ambient lighting (Table 4). Scrolling speed and this time also loading speed were still prominent topics. The comments specified that especially loading of the volumetric data sets (CT and MR images) was perceived as slow.

Cognitive Task Characteristics

Eighty percent of the participants agreed that digital radiology progress testing corresponds better to clinical practice than paper-based testing. Again, participants found that digital image questions corresponded significantly better to clinical practice than paper-based image questions [mean (SD) = 3.4 (1.0, n = 110) and mean (SD) = 3.1 (1.0, n = 110), respectively, $t(109) = 3.28$, $P < 0.01$, $d = 0.30$]. Apart from the positive comments toward digital testing with volumetric images, some participants feel that image reading should rather be tested in clinical practice, whereas the progress test should focus on knowledge and interpretation of static images. One of the comments: *"I think the best clinical practice test is real clinical practice. The progress test is primarily a knowledge test and navigating through images does not add much."*

DISCUSSION

We described the development and implementation process of a digital simulation-based assessment in radiology residency. We aimed to improve the authenticity of image interpretation assessment. However, the authenticity of the first DT was not rated higher than the PTs. After optimizing test characteristics, image manipulation options, and environmental conditions, participants favored the authenticity of the digital image questions over paper-based image questions.

An improved representation of clinical practice with volumetric images was reported previously among medical students,[22] who have very limited experience with interpreting images in clinical practice. Standards for the viewing task are probably higher for residents, who are used to working with high-quality images and advanced, high-speed image manipulation equipment. Unfortunately, computers available for large-scale tests usually do not meet the high criteria of screen size, screen resolution, and processor speed used in radiology practice. In our study, this discrepancy was reflected by the initial high number of suggestions for improvement of the viewing task, expressed by the residents.

One of the most important challenges was how to improve perceived image quality. Digital image quality was initially rated lower than paper-based image quality. Based on the comments, this was probably due to the small displays with limited screen resolution that were available in this large-scale assessment setting. The screen resolution was less than the 1280×1024 displays that were previously recommended for radiology assessments.[20] Small screen size had to be compensated by optimizing image manipulation options. We therefore introduced and optimized the zooming function and introduced a full screen option for cases with multiple images. Besides, speed and smoothness of scrolling and zooming were criticized by the participants and had to be optimized. Our results underscore that only implementing image manipulation options is not enough, and an optimal functionality of the options is crucial to reach a satisfactory simulation. This relates to the human-computer interaction literature, showing that intuitive and direct interactions are crucial for user acceptance of a computerized system.[32]

Optimizing the ambient light was another challenge. A high level of ambient light can have a negative effect on observer performance.[33–35] Some studies have reported that the issue of ambient light can be partially compensated for by means of interactive contrast adjustments.[33,36] Even though participants could adjust contrast settings, a substantial portion of their comments was related to hindrance of ambient light. Because there was no possibility to dim lights, we ultimately turned the lights off, which was appreciated by the participants. On the other hand, there is evidence that a dark reading room may increase visual fatigue due to pupil contraction and dilation, which negatively affects reader performance.[37] It is therefore recommended to include possibilities for light control in assessment rooms for image interpretation tests.

Screen size, image manipulation options, and the level of ambient light seem to be crucial factors for establishing an optimal reflection of clinical practice in radiology assessment. The importance of these factors probably varies in other visual domains, such as dermatology and clinical pathology. For example, image manipulation options are probably more important in clinical pathology[38] than in clinical dermatology, whereas ambient light reflection on computer screens may affect any image interpretation task.

Another important consideration is the increased time needed to review volumetric images. Loading of and scrolling through volumetric images take time. Comments regarding low loading and scrolling speed increased in DT 2 and 3, probably because of a higher number of volumetric images. The improved reflection of clinical practice with volumetric images has a trade-off with time constraints. Increased time and effort of participants, faculty, and staff support are common drawbacks of simulations.[39] It requires careful consideration whether the desired test validity justifies the investments.[40] We therefore recommend that volumetric images should only be used if they are considered to better match the goal of the question. For

example, a question that aims to test the ability to distinguish a tumor from an infection may not necessarily require a volumetric image data set. Volumetric image data sets are necessary to test whether trainees are able to detect abnormalities (perception skills) on a CT or MR image or to test whether they can recognize a pattern of abnormalities.

Some participants commented that image interpretation can be better tested in actual clinical practice. Why put all these efforts in a simulation-based test if we can test image interpretation in clinical practice? There are some important disadvantages of testing image interpretation skills in real clinical practice. Because time and resources are limited, a supervisor may not always have time to check the findings and the interpretation of a resident. Furthermore, judgments between supervisors may differ. Besides, every trainee encounters different patient cases, with variable content and difficulty levels. These drawbacks are serious threats for test quality. Our standardized simulation-based test does not have these disadvantages. It has already proven to be a valid and reliable test,[24] and we improved its authenticity. In our view, these advantages in test quality justify the time and effort that are needed to develop the questions and to administer the test.

Throughout the development process, we introduced multiple choice, hotspot, and long-menu questions to better reflect the cognitive task of image interpretation. In most of the tests, the multiple choice, hotspot, and long-menu questions reached an average r_{it} value of 0.30 or larger, as is recommend for high-stake tests,[41] in contrast to the true/false questions that showed average r_{it} values from 0.15 to 0.22. However, the long-menu questions have some disadvantages for teachers, because they require extensive lists of answer options, including synonyms, and are time-consuming to construct.[12]

The development process with changes in image manipulation options, ambient lighting, time limits, and question types resulted in an improved authenticity of the digital simulation-based assessment compared with its paper-based counterpart. Because we implemented and developed the test in three phases and changed multiple factors in each phase, we cannot determine the effect of each separate factor. We can only conclude that the cumulative set of changes was advantageous and recommend that these factors be carefully considered when developing an image interpretation test.

Some limitations should be addressed. Although the entire cohort of Dutch radiology residents participated in the three DTs, we cannot generalize these results to radiology programs in other countries. In addition, response rates of the questionnaires were moderate and slightly decreased throughout the study. However, only the fifth year residents were slightly underrepresented in the questionnaire responses, possibly because they felt that they would not benefit from future test improvements.

When interpreting the test results over time, we should acknowledge differences in group composition across the three tests, primarily due to residents continuously phasing in and out of the program throughout the year. However, the average training year remained virtually constant across the tests.

Although radiology residents do extensively work with volumetric images during their residency, they do not have as much experience as radiologists have. Especially the residents who just started their residency may not be the best evaluators of authenticity, and radiologists may have a different perspective. However, it is important that the test takers underscore the authenticity of the test for the sake of test acceptation and to stimulate learning.

CONCLUSIONS

This study underscores that authenticity of a simulation does not necessarily increase with higher fidelity. Simulating the image interpretation task of radiology practice in a large-scale assessment setting is challenging, because of technological limitations. Optimizing image manipulation options, the level of ambient light, time limits, and question types can help improve authenticity of simulation-based radiology assessments. In our view, the improved test quality of simulation-based radiology assessments justifies the required time and effort for its development.

REFERENCES

1. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;65:S63–S67.
2. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA* 2009;302:1316–1326.
3. Cook DA, Triola MM. Virtual patients: a critical literature review and proposed next steps. *Med Educ* 2009;43:303–311.
4. Okuda Y, Bryson EO, DeMaria S Jr, et al. The utility of simulation in medical education: what is the evidence? *Mt Sinai J Med* 2009;76:330–343.
5. Bland AJ, Topping A, Tobbell J. Time to unravel the conceptual confusion of authenticity and fidelity and their contribution to learning within simulation-based nurse education. A discussion paper. *Nurse Educ Today* 2014;34(7):1112–1118.
6. Gulikers JTM, Bastiaens TJ, Kirschner PA. A five-dimensional framework for authentic assessment. *Educ Technol Res Dev* 2004;52:67–86.
7. Andriole KP, Wolfe JM, Khorasani R, et al. Optimizing analysis, visualization, and navigation of large image data sets: one 5000-section CT scan can ruin your whole day. *Radiology* 2011;259:346–362.
8. Reiner BI, Siegel EL, Siddiqui K. Evolution, of the digital revolution: a radiologist perspective. *J Digit Imaging* 2003;16:324–330.
9. van der Gijp A, van der Schaaf MF, van der Schaaf IC, et al. Interpretation of radiological images: towards a framework of knowledge and skills. *Adv Health Sci Educ Theory Pract* 2014;19(4):565–580.
10. van der Gijp A, Ravesloot CJ, van der Schaaf MF, et al. Volumetric and two-dimensional image interpretation show different cognitive processes in learners. *Acad Radiol* 2015;22(5):632–639.
11. Gulikers JTM, Bastiaens TJ, Kirschner PA, et al. Authenticity is in the eye of the beholder: student and teacher perceptions of assessment authenticity. *J Vocat Educ Train* 2008;60(4):401–412.
12. van Bruggen L, Manrique-van Woudenbergh M, Spierenburg E, et al. Preferred question types for computer-based assessment of clinical reasoning: a literature study. *Perspect Med Educ* 2012;1:162–171.
13. Zafar S, Safdar S, Zafar AN. Evaluation of use of e-learning in undergraduate radiology education: a review. *Eur J Radiol* 2014;83:2277–2287.
14. den Harder AM, Frijlingh M, Ravesloot CJ, et al. The importance of human-computer interaction in radiology e-learning. *J Digit Imaging* 2016;29(2):195–205.
15. Ernst RD, Sarai P, Nishino T, et al. Transition from film to electronic media in the first-year medical school gross anatomy lab. *J Digit Imaging* 2003;16:337–340.
16. Howlett D, Vincent T, Watson G, et al. Blending Online Techniques with Traditional Face to Face Teaching Methods to Deliver Final Year Undergraduate Radiology Learning Content. *Eur J Radiol* 2011;78:334–341.

17. Turmezei TD, Tam MD, Loughna S. A survey of medical students on the impact of a new digital imaging library in the dissection room. *Clin Anat* 2009;22:761–769.

18. Petersson H, Sinkvist D, Wang C, et al. Web-based interactive 3D visualization as a tool for improved anatomy learning. *Anat Sci Educ* 2009;2:61–68.

19. Rengier F, Hafner MF, Unterhinninghofen R, 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;82:1366–1371.

20. Krupinski EA, Becker GJ, Laszakovits D, et al. Evaluation of off-the-shelf displays for use in the American Board of Radiology maintenance of certification examination. *Radiology* 2009;250:658–664.

21. Mullins ME, Will M, Mehta A, et al. Evaluating medical students on radiology clerkships in a filmless environment: use of an electronic test prepared from PACS and digital teaching collection images. *Acad Radiol* 2001;8:514–519.

22. Ravesloot CJ, van der Schaaf MF, van Schaik JP, et al. Volumetric CT-images improve testing of radiological image interpretation skills. *Eur J Radiol* 2015;84(5):856–861.

23. Albanese M, Case SM. Progress testing: critical analysis and suggested practices. *Adv Health Sci Educ Theory Pract* 2016; 21(1):221–234.

24. Ravesloot C, van der Schaaf M, Haaring C, et al. Construct validation of progress testing to measure knowledge and visual skills in radiology. *Med Teach* 2012;34:1047–1055.

25. Morita J, Miwa K, Kitasaka T, et al. Interactions of perceptual and conceptual processing: expertise in medical image diagnosis. *Int J Hum Comput Stud* 2008;66:370–390.

26. Norman GR, Coblentz CL, Brooks LR, et al. Expertise in visual diagnosis: a review of the literature. *Acad Med* 1992;67:S78–S83.

27. Kundel HL, Nodine CF, Carmody D. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Invest Radiol* 1978;13:175–181.

28. Donald JJ, Barnard SA. Common patterns in 558 diagnostic radiology errors. *J Med Imaging Radiat Oncol* 2012;56:173–178.

29. Ebel RL. Estimation of the reliability of ratings. *Psychometrika* 1951;16:407–424.

30. Ravesloot CJ, Van der Schaaf MF, Muijtjens AM, et al. The don't know option in progress testing. *Adv Health Sci Educ Theory Pract* 2015;20:1325–1338.

31. Ravesloot CJ, van der Gijp A, van der Schaaf MF, et al. Support for external validity of radiological anatomy tests using volumetric images. *Acad Radiol* 2015;22:640–645.

32. O'Brien MA, Rogers WA, Fisk AD. Developing a framework for intuitive human-computer interaction. *Proc Hum Factors Ergon Soc Annu Meet* 2008;52:1645–1649.

33. Fuchsjager MH, Schaefer-Prokop CM, Eisenhuber E, et al. Impact of ambient light and window settings on the detectability of catheters on soft-copy display of chest radiographs at bedside. *AJR Am J Roentgenol* 2003;181:1415–1421.

34. Hellén-Halme K, Lith A. Effect of ambient light level at the monitor surface on digital radiographic evaluation of approximal carious lesions: an in vitro study. *Dentomaxillofac Radiol* 2012;41(3):192–196.

35. Uffmann M, Prokop M, Kupper W, et al. Soft-copy reading of digital chest radiographs: effect of ambient light and automatic optimization of monitor luminance. *Invest Radiol* 2005;40(3):180–185.

36. Goo JM, Choi JY, Im JG, et al. Effect of monitor luminance and ambient light on observer performance in soft-copy reading of digital chest radiographs. *Radiology* 2004;232:762–766.

37. Pollard BJ, Chawla AS, Delong DM, et al. Object detectability at increased ambient lighting conditions. *Med Phys* 2008;35:2204–2213.

38. Jaarsma T, Jarodzka H, Nap M, et al. Expertise in clinical pathology: combining the visual and cognitive perspective. *Adv Health Sci Educ Theory Pract* 2015;20:1089–1106.

39. Srinivasan M, Hwang JC, West D, et al. Assessment of clinical skills using simulator technologies. *Acad Psychiatry* 2006;30:505–515.

40. Issenberg SB, McGaghie WC, Hart IR, et al. Simulation technology for health care professional skills training and assessment. *JAMA* 1999;282:861–866.

41. Downing SM, Yudkowsky R. *Statistics of Testing. Assessment in Health Professions Education*. New York: Routledge; 2009.