Support for External Validity of Radiological Anatomy Tests Using Volumetric Images

Cécile J. Ravesloot, MD, Anouk van der Gijp, MD, Marieke F. van der Schaaf, PhD, Josephine C. B. M. Huige, MD, Koen L. Vincken, PhD, Christian P. Mol, MSc, Ronald L. A. W. Bleys, MD, PhD, Olle T. ten Cate, PhD, Jan P. J. van Schaik, MD, PhD

Rationale and Objectives: Radiology practice has become increasingly based on volumetric images (VIs), but tests in medical education still mainly involve two-dimensional (2D) images. We created a novel, digital, VI test and hypothesized that scores on this test would better reflect radiological anatomy skills than scores on a traditional 2D image test. To evaluate external validity we correlated VI and 2D image test scores with anatomy cadaver–based test scores.

Materials and Methods: In 2012, 246 medical students completed one of two comparable versions (A and B) of a digital radiology test, each containing 20 2D image and 20 VI questions. Thirty-three of these participants also took a human cadaver anatomy test. Mean scores and reliabilities of the 2D image and VI subtests were compared and correlated with human cadaver anatomy test scores. Participants received a questionnaire about perceived representativeness and difficulty of the radiology test.

Results: Human cadaver test scores were not correlated with 2D image scores, but significantly correlated with VI scores ($r = 0.44$, $P < .05$). Cronbach's α reliability was 0.49 (A) and 0.65 (B) for the 2D image subtests and 0.65 (A) and 0.71 (B) for VI subtests. Mean VI scores (74.4%, standard deviation 2.9) were significantly lower than 2D image scores (83.8%, standard deviation 2.4) in version A ($P < .001$). VI questions were considered more representative of clinical practice and education than 2D image questions and less difficult (both $P < .001$).

Conclusions: VI tests show higher reliability, a significant correlation with human cadaver test scores, and are considered more representative for clinical practice than tests with 2D images.

Key Words: Testing; radiology education; volumetric datasets; volumetric images; radiological image interpretation.

Radiologists and other medical specialists involved in interpreting radiological images are confronted with large datasets and ample options for image manipulation (1). Although radiologists used to view cross-sectional images as single slices presented next to each other (ie, tile viewing), nowadays, the use of innovative image displaying software is the norm. This allows the radiologist to scroll through three-dimensional (3D) datasets (stack viewing), adjust window level, and use advanced image reconstruction tools, such as on the fly multiplanar reformatting. The data for one cross-sectional patient investigation involve a volumetric image (VI) containing up to hundreds of slices, which can be scrolled through in various planes and contrast settings. A vast amount of visual information must be processed and interpreted by the observer (2). Radiological image interpretation has changed significantly and consequently requires different skills (1–4). It is therefore important that radiology education should change accordingly (5).

Acquiring basic radiological knowledge and image interpretation skills for medical students is increasingly important, as diagnostic imaging has become a prominent diagnostic tool in daily clinical practice (5,6). Specifically, the knowledge of radiological anatomy is required for medical doctors of various specialisms to recognize abnormalities on radiological images and to understand the radiology report (7,8). Efforts are made to innovate and digitalize radiology education; however, the contents of these curricula vary and are often not supported by empirical evidence (9,10). In particular, studies on the

Acad Radiol 2015; 22:640–645

From the Department of Radiology, University Medical Center Utrecht, Room E01.132, Heidelberglaan 100, 3508 GA, Utrecht, The Netherlands (C.J.R., A.V.G, J.C.B.M.H., J.P.J.S.); Department of Education, Utrecht University, Utrecht, The Netherlands (M.F.V.S.); Image Sciences Institute, University Medical Center Utrecht, Utrecht, The Netherlands (K.L.V., C.P.M.); Department of Anatomy, University Medical Center Utrecht, Utrecht, The Netherlands (R.L.A.W.B.); and Center for Research and Development of Education, University Medical Center Utrecht, Utrecht, The Netherlands (O.T.C.). Received August 12, 2014; accepted December 11, 2014. Funding source: This work was partly financially supported by the SURF Foundation, Collaborative Organization for ICT in Dutch higher education and research. SURF had no involvement in the study design, analysis, interpretation of the data, or drafting of the article. **Address correspondence to:** C.J.R. e-mail: C.J.Ravesloot@umcutrecht.nl

http://dx.doi.org/10.1016/j.acra.2014.12.013

development of high quality radiology tests are scarce. Furthermore, most radiology tests do not do justice to the major developments in radiological image interpretation practice. For example, most radiology tests or self-assessment tools do not contain VIs or allow for image manipulation (2D image test) (11). Pass or fail decisions in traditional radiology tests might therefore become increasingly meaningless given they may reflect measures of irrelevant competence. High quality radiology tests are consequently essential to ensure adequate levels of radiological performance among medical doctors.

To argue a high test quality, evidence for reliability and support for validity of the test needs to be gathered (12). Reliability refers to the accuracy and reproducibility of test scores. Validity implies that the test measures what it is intended to measure, and that therefore decisions regarding students' skills based on their scores are valid. More authentic tests, reflecting clinical practice, contribute to validity, because the skills assessed are in accordance with those used in practice (13). Almost all current radiology tests are based on 2D images, that is, a single slice is taken from a VI, either based on a computed tomography (CT) or magnetic resonance (MR) scan. The validity of such tests might be at stake, as arguably these 2D image tests do not measure the intended radiological skills needed in the altered radiological practice. Digitalization and introducing VI in radiology tests might improve test validity by increasing its representativeness of clinical practice. The first results from radiology tests with VI are promising and indicate that reliability and perceived representativeness for clinical practice are higher for VI tests than for traditional 2D image tests (14). Additionally, students considered VI tests to better reflect image interpretation skills required in clinical practice than 2D image tests (14). The external validity of a test is another useful objective measure of its validity. External validity addresses whether test scores correlate to other measurements of the same knowledge and skills intended to be tested (12,15).

In this study, we aimed to gather evidence for external validity of VI testing in radiological anatomy education of medical students. We correlated VI test scores to human cadaver anatomy test results as an external measure of knowledge on 3D aspects of anatomy, and compared the results to the correlation of 2D image test scores to this measure. A golden standard for radiological anatomic skill performance is not available; therefore, we assumed that a human cadaver anatomy test would serve as a good alternative, approximating radiological anatomy interpretation skills. We hypothesized that the understanding of 3D anatomy is better resembled by VI interpretation than by the interpretation of 2D images. In addition, we evaluated indications of reliability, perceived representativeness of clinical practice, and difficulty of 2D image versus VI questions in radiology.

MATERIALS AND METHODS

Study Design

In April 2012, 278 medical students at University Utrecht took a digital radiology test with 2D image and VI questions at the end of their second preclinical year. Written informed consent was provided by 246 students before the test commenced. After the test, students received a digital questionnaire to measure both perceived representativeness of clinical practice and radiology education as well as perceived difficulty of 2D image and VI questions. All study participants were invited to take a traditional human cadaver anatomy test. Thirty-three students agreed to participate and took the human cadaver test 2 months after the radiology test. Again, written informed consent was provided before the test. Anonymous questionnaire responses and test scores were analyzed to evaluate 2D image and VI test quality. Ethical approval was obtained from the Ethical Review Board of the Netherlands Association for Medical Education.

Population

All participants had completed a 2-year radiology education program including basic radiological skills on prevalent diseases and radiological anatomy as part of their preclinical medical training. They attended 12 2-hour case–based small group classes consisting of 8–10 students, in which they practiced radiological chest and abdominal anatomy with 2D and volumetric CT scans among other things. Groups were instructed by senior medical students and supervised by radiology residents (16). Approximately 100 hours of study time in the medical curriculum was devoted to radiology. All participants had studied gross anatomy in human cadavers as part of the regular medical curriculum.

Instrumentation

Radiology test. The test consisted of 75 questions, including 40 CT-anatomy questions. The remaining 35 questions concerned basic radiological image interpretation skills and knowledge of prevalent diseases. Twenty CT-anatomy questions involved a whole volumetric dataset of either a normal abdominal or a chest CT scan (VI questions). In the remaining 20 CT-anatomy questions, each question concerned one slice selected from one of these two CT scans (2D image questions). Half of both the 2D image and the VI questions were phrased as, for example, "See normal CT-scan. Mark *the aorta ascendens*." We called these questions "indication questions." To answer an indication question participants had to put a marker in the requested anatomic structure in the image (2D image or VI), see Figure 1. The other half of the questions was phrased as "See normal CT-scan. Which anatomic structure is marked red? Be as specific as possible." We called this question type "identification questions." To answer an identification question participants had to choose the right answer from a list with options containing up to 2000 anatomic structures. Participants could search in the option list by typing at least two letters of their answer in the drop down box. A complete overview of all questions used in the tests is provided in the digital Supplement. All participants started with the 2D

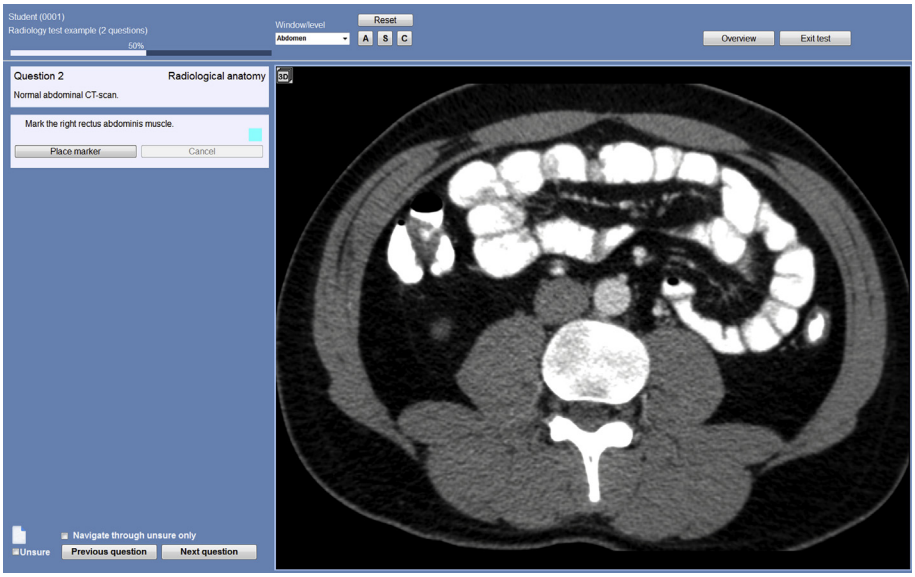

Figure 1. Example volumetric image question. Participants could change viewing direction, by pushing buttons A (axial), S (sagittal), and C (coronal), and alter contrast setting (abdomen, bone, and lung setting) by selecting a preset window level (menu below "Window/level").

image questions, immediately followed by the VI questions. For logistical reasons students were tested in two groups at different time-points (two dates in April 2012) and both groups completed a different test version.

The score model for each question was discussed and derived by the examiners before the test. All answers were checked automatically by the testing program (15) based on this predefined score model. For each indication question the area containing the requested anatomic structure was marked in the image by the examiners (3D region in case of VIs and 2D region for 2D images). If the marker was placed within this area the answer was considered correct and the participant assigned 1 point, if the marker was not placed within the marked area no points were assigned. On the basis of the score model, identification questions were considered correct (1 point), partly correct (0.25, 0.5, or 0.75 points), or incorrect (zero points). Each participant received a 2D image and a VI test percentage score; a score of 20 points resulted in a percentage score of 100% for each subtest.

Testing program. The test was administered in (15), a software package designed for testing with volumetric datasets. Participants were able to adjust window settings in both 2D image and VI, and viewed VI in stack mode with the ability to change viewing direction (Fig 1) (15).

Human cadaver anatomy test. During a traditional human cadaver test students were asked to identify or indicate 40 anatomic structures in prosected human bodies under the supervision of an anatomy teacher. Phrasing of the questions and tested anatomic structures were identical to those in the radiology test (see digital Supplement). Two questions were slightly adapted, however, because they included an anatomic structure not suitable for identification in a human cadaver. For example, "gastric bubble" in the radiology test was changed to "gastric fundus" in the human cadaver test. Participants who took version A of the radiology test took the human cadaver test with the anatomic structures of version B of the radiology test, and *vice versa*.

Scores were calculated based on a scoring model derived by the examiners group. Every participant received a percentage human cadaver test score, in which a score of 40 was equal to a 100% percent score.

Questionnaire radiology test. The questionnaire concerning the radiology test included four questions on perceived representativeness of clinical practice and education, and 10 questions on difficulty of the 2D image and VI questions. This questionnaire was completed by 130 participants.

Analysis

Radiology 2D image and VI subtest scores were correlated with human cadaver test scores using the Pearson correlation coefficient. Mean radiology test scores (sum of 2D image and VI subtest scores) of cadaver test participants were compared to mean scores of radiology test participants who did not volunteer in the cadaver test using *t* tests. Reliabilities estimated with Cronbach's α values of the 2D image and VI subtest scores were compared. Predicted Cronbach's α values of 2D image and VI test scores including 40 questions were calculated with the Spearman Brown formula (17). Questionnaire responses to questions on perceived representativeness for 2D image and VI questions were compared using paired *t* tests. Means of five questionnaire items measuring difficulty on different aspects were calculated for 2D image and VI questions and compared using paired *t* tests. The estimated reliabilities of these two difficulty scales were Cronbach's α 0.87 (2D image questions) and 0.89 (VI questions), respectively.

TABLE 1. Radiology Test Results, Baseline Characteristics, Including Mean Test Scores and Standard Deviations (SDs), Cronbach's α as Reliability Estimates of Two-Dimensional (2D) Image and Volumetric Image (VI) Test Scores (20 and 40 Questions) per Test Version

Test characteristics	Version A	Version B
Number of participants	121	125
2D image questions (k = 20)		
α	0.49	0.65
α predicted for k = 40*	0.66	0.79
Mean score % (SD)	78.8 (10.8)	83.8† (11.8)
VI questions (k = 20)		
α	0.65	0.71
α predicted for k = 40*	0.79	0.83
Mean score % (SD)	78.8 (13.0)	74.7† (14.4)

k = Number of questions.
*P < .001.
†Cronbachs α predicted for test scores including 40 questions.

TABLE 2. Human Cadaver Test Results, and Correlations with Two-Dimensional (2D) Image and Volumetric Image (VI) Radiology Test Scores

Test characteristics	Version A (k = 40)	Version B (k = 40)
Number of participants	19	14
α	0.62	0.81
Mean score (standard deviation)	60.3 (2.5)	73.0 (3.5)
Correlation (Pearson) with radiology subtest scores		
2D image questions	0.07	
VI questions	0.44*	

k = Number of questions.
*Significantly different P < .05.

RESULTS

Radiology Test

Mean scores and Cronbach's α values of versions A and B of the radiology test are shown in Table 1. Mean 2D image scores were significantly higher than VI scores for test B participants (t (124) = 8.7; P < .001, eta squared = 0.36). Mean radiology test scores of human cadaver test participants (83.2%, standard deviation [SD] 8.0) were higher than for radiology test participants who did not volunteer in the human cadaver test (78.2%, SD 11.5). This difference in mean radiology test scores was significant for test A participants (t (119) = −2.4; P = .02, eta squared = 0.04).

Correlation Radiology and Human Cadaver Test Scores

Nineteen of the radiology test version A participants and 14 of the radiology test version B participants agreed to take the human cadaver test. Mean human cadaver test scores and Cronbach's α values are shown in Table 2. Radiological VI subtest scores had a significant correlation with human cadaver test scores, whereas radiology 2D image subtest scores did not correlate with human cadaver test scores.

Questionnaire Responses

The response rate was 54%. Respondents considered VI questions significantly more representative of radiology education and clinical practice compared to 2D image questions (see Table 3). 2D image questions were considered significantly more difficult than VI questions.

DISCUSSION

This study shows that VI contributes to the external validity and reliability of radiological anatomy testing. Contrary to 2D image questions, VI questions on radiological anatomy correlated significantly to an external measure of 3D anatomy knowledge. Accordingly, digital radiological VI interpretation in current clinical practice requires a more holistic 3D understanding of anatomy, which is better reflected by VI than 2D image radiology test scores. To our knowledge, no earlier studies have investigated the correlation between radiology test scores and human cadaver test scores. Previous studies have found that teaching anatomic knowledge is important for radiology image interpretation; for example, that 2D CT anatomy interpretation can be improved by learning sectional anatomy (18). Our study results suggest that 3D, in contrast to sectional, anatomic knowledge is particularly important for current radiological anatomic image interpretation.

Consistent with results on a previous retrospective study, VI improved test quality on two other aspects: reliability and perceived representativeness for clinical practice (14). Increased reliability means that fewer VI questions are needed to obtain accurate and reproducible test results than in 2D image tests. A possible explanation for this difference in reliability between VI and 2D image questions could be a wider dispersion of test scores for VI questions, which is supported by larger SDs. More advanced knowledge and skills are probably required to answer a VI question than a 2D image question. This enhances the difference between students by improving discrimination between high and low performance, which adds to reliability.

Student performance on VI questions was lower than on 2D image questions, which corroborates the previous retrospective comparison (14). A possible explanation is that VI questions require more extensive searching (eg, scrolling and changing views), whereas in a 2D image question the anatomic structure of interest is always shown. In addition, VIs also provide much more information and the answer possibilities are rather exhaustive. Remarkably, VI questions were perceived as less difficult. This phenomenon has been found before and points to a false sense of security, from the availability of information in a task, for example, in open book examinations (14,19).

The 2D image question scores showed a lack of correlation with an external measure, lower reliability, better student performance, and smaller dispersion. One explanation for this

TABLE 3. Means (*M*) and Standard Deviations (SDs) of Responses on Questionnaire Items Concerning Perceived Representativeness and Scale of Difficulty Items of Two-Dimensional (2D) Image and Volumetric Image (VI) Questions

Questionnaire items	*M*		Number of Responses
	2D Image	VI	
Perceived representativeness			
Scale: 1–5 ("completely disagree" to "completely agree")			
2D image/VI questions reflect radiology education*	3.2 (1.2)	4.2 (0.9)	130
2D image/VI questions reflect clinical practice*	3.2 (1.2)	4.3 (0.8)	129
Difficulty			
Scale: 1–5 ("very easy" to "very difficult")			
Mean scale of items on difficulty of 2D image/VI questions*	3.5 (0.7)	2.7 (0.7)	131

*Significantly different at $P < .001$.

may be that in the 2D image questions skills and knowledge of radiological anatomy were underrepresented. Such underrepresentation is a threat to validity and implies that 2D image test results might not reflect students' performance on radiological image interpretation (20).

Several limitations of this study should be noted. First, one could question the human cadaver test as an external measure for radiological anatomy interpretation skills. However, as a golden standard does not exist, which is usually the case for tests, we decided on a test that best reflects 3D anatomy knowledge and visual skills. Second, only a small, nonrandom sample of the study participants took the human cadaver test. Mean radiology test scores of the human cadaver test participants were slightly higher than the mean scores of the rest of the radiology test participants. However, these differences in scores were only significant in one of the two testing groups (A), and the effect size was low to moderate, indicating only a small selection bias favoring high performers. Third, the results about perceived representativeness for clinical practice should be interpreted with some caution, as students generally do not have clinical experience and cannot compare the test to real clinic practice. This finding should therefore be verified with a population experienced in clinical practice. Fourth, we only investigated radiological anatomy tests; however, testing radiological image interpretation of abnormal cases could also benefit from the use of VI and should be subject of future research. Fifth, reliabilities of all four subtests were relatively low, because there were only 20 questions per subtest. However, predicted reliabilities for test scores based on 40 questions using the Spearman Brown formula were sufficient to high. The difference in reliability between the four subtests might be because of differences in asked anatomic structures, but also owing to a larger dispersion in both 2D image and VI test scores for test B participants compared to test A participants (see larger SDs in group B in Table 1).

The results indicate that VIs improve radiological anatomy test quality on all studied quality aspects. The increase of radiological images in many medical specialties asks for a high test quality in medical education to warrant a high level of radiological image interpretation skills of our future medical doctors. VIs with stack viewing and multiplanar reformatting tools should therefore be included in radiology tests, especially in radiological anatomy testing.

ACKNOWLEDGMENTS

We thank Bobby G and Suzannah Stuijfzand of the School of Experimental Psychology, University of Bristol, UK, and the Department of Psychology, University of Reading, UK, respectively, for reviewing and editing the manuscript and refining the use of English language.

SUPPLEMENTARY DATA

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.acra.2014.12.013.

REFERENCES

1. 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.
2. Reiner BI, Siegel EL, Siddiqui K. Evolution of the digital revolution: a radiologist perspective. J Digit Imaging 2003; 16:324–330.
3. Wang XH, Durick JE, Lu A, et al. Characterization of radiologists' search strategies for lung nodule detection: slice-based versus volumetric displays. J Digit Imaging 2008; 21(suppl 1):S39–S49.
4. Venjakob A, Marnitz T, Mahler J, et al. Radiologists' eye gaze when reading cranial CT images. SPIE Proceedings, volume 8318, Medical Imaging 2012; 83180B.
5. Pascual TNB, Chhem R, Wang SC, et al. Undergraduate radiology education in the era of dynamism in medical curriculum: an educational perspective. Eur J Radiol 2011; 78:319–325.
6. Gunderman RB, Siddiqui AR, Heitkamp DE, et al. The vital role of radiology in the medical school curriculum. AJR Am J Roentgenol 2003; 180:1239–1242.
7. 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; 19:565–580.
8. Erkonen WE, Albanese MA, Smith WL, et al. Effectiveness of teaching radiologic image interpretation in gross anatomy. A long-term follow-up. Invest Radiol 1992; 27:264–266.
9. Kourdioukova EV, Valcke M, Derese A, et al. Analysis of radiology education in undergraduate medical doctors training in Europe. Eur J Radiol 2011; 78:309–318.
10. Lee JS, Aldrich JE, Eftekhari A, et al. Implementation of a new undergraduate radiology curriculum: experience at the University of British Columbia. Can Assoc Radiol J 2007; 58:272–278.

11. Scarsbrook AF, Graham RNJ, Perriss RW. Radiology education: a glimpse into the future. Clin Radiol 2006; 61:640–648.
12. Messick S. Validity. In: Linn LR, ed. Educational measurement. 3rd ed. 1989; 13–103.
13. van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. Med Educ 2005; 39:309–317.
14. Ravesloot CJ, Van der Schaaf MF, Van Schaik JPJ, et al. Volumetric Images improve the testing of Radiological Image Interpretation Skills. Eur J of Radiol 2015; http://dx.doi.org/10.1016/j.ejrad.2014.12.015.
15. Poldner E, Simons PRJ, Wijngaards G, et al. Quantitative content analysis procedures to analyse students reflective essays: a methodological review of psychometric and edumetric aspects. Educ Res Rev 2012; 7:19–37.
16. van den Berk IA, van de Ridder JM, van Schaik JP. Radiology as part of an objective structured clinical examination on clinical skills. Eur J Radiol 2011; 78:363–367.
17. Ebel RL. Measuring educational achievement. New Jersey: Prentice-Hall; 1965.
18. Barros de N, Rodrigues CJ, Rodrigues AJ, Jr, et al. The value of teaching sectional anatomy to improve CT scan interpretation. Clin Anat 2001; 14:36–41.
19. Dale VH, Wieland B, Pirkelbauer B, et al. Value and benefits of open-book examinations as assessment for deep learning in a post-graduate animal health course. J Vet Med Educ 2009; 36:403–410.
20. Messick S. Validity of psychological assessment: validation of inferences from Persons' responses and performances as scientific inquiry into score meaning. Am Psychol 1995; 50:741–749.