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