



Direct Observation of Procedural Skills in Radiology

Vaqaar Bari¹

OBJECTIVE. The purpose of this perspective is to describe the reliability and feasibility of methods such as direct observation of procedural skills and multisource feedback in assessment of the performance of radiology residents.

CONCLUSION. Workplace-based assessments such as direct observation of procedural skills have a role in the formative assessment of radiology residents. They can be used to evaluate residents' performance, provide feedback, and identify areas for improving performance and filling in identified gaps.

Problem Statement and Context: Postgraduate medical education and training is a dynamic field undergoing major changes around the world. The old method of learning through apprenticeship with one or more senior clinical colleagues over long working hours and observing large numbers of healthy and ill or injured patients is being challenged. Because of limits on the hours that can be worked and shorter training, as much time as possible at work must be used for learning [1].

The number of residents at the radiology department at my university hospital has increased from four to six per year, and residency duration has increased from 4 to 5 years, increasing the need for an appropriate system for evaluating learning. The problem of inadequate or lack of observation of residents during the training period has been recognized. The radiology department performs a large number of imaging-guided procedures every year. On an average day, at least seven ultrasound-guided procedures are performed. In a week's rotation, residents are given an opportunity to perform as many procedures as possible according to their level of residency training. Continual assessment, both formative and summative, is needed to meet the objective of ensuring clinical competence. Direct observation of procedural skills is a method that can have a role in the formative assessment of radiology residents. It can be used to evaluate residents' performance, provide feedback, and identify areas for improving performance and filling in identified gaps.

Miller's pyramid (Fig. 1) is a useful way of describing levels of competence. The steps progress from knows, which reflects applied knowledge, through knows how, which requires more than knowledge alone, and shows how, which requires an ability to show clinical competency, to does. Performance-based evaluation is the assessment of does in Miller's pyramid. Direct observation of procedural skills, developed by the Royal College of Physicians in the United Kingdom, requires an assessor to directly observe a trainee undertaking a procedure and then grade the performance of specific predetermined components of the procedure. In addition to the procedure itself, these skills include communication and the informed consent process.

Literature Review

Radiology differs from other specialties in many ways. Trainees are protected in their early years by working in a close apprenticeship with their supervisors. Their knowledge and skills in the workplace are assessed daily but not in a standardized way, and the findings are not formally documented.

Performance-based methods such as direct observation of procedural skills are ideal in the assessment of diagnostic and interventional radiologic procedures [2]. Virtual reality simulator models can help in training and assessment of the core skills of interventional radiology and reduce the time required for achieving and maintaining competence. A number of vascular interventional radiology visual simula-

Keywords: assessment, direct observation of procedural skills, performance

DOI:10.2214/AJR.09.4068

Received December 4, 2009; accepted without revision December 29, 2009.

¹Radiology Department, Aga Khan University Hospital, Stadium Rd., PO Box 3500, Karachi, 74800 Pakistan. Address correspondence to V. Bari (vaqaar.bari@aku.edu).

WEB

This is a Web exclusive article.

AJR 2010; 195:W14–W18

0361–803X/10/1951–W14

© American Roentgen Ray Society

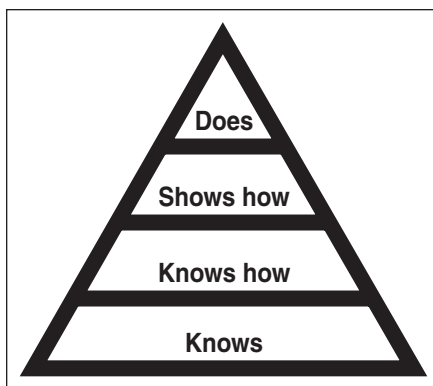


Fig. 1—Miller's pyramid.

tors and some visceral interventional radiology simulators have been developed, but none has been validated. If proved reliable and valid, these simulators can be used in task analysis. Training sessions have been conducted in various workshops and radiology departments whereby an animal liver has been used for training residents in ultrasound-guided needle localization. This training is valuable for developing competence before the trainee performs a procedure on an actual patient.

Some skills, mainly visual skills related to orientation and spatial negotiation, can be taught in models, as in surgery. This method, however, has limitations in interventional radiology, which relies heavily on the sense of touch. Both patients and trainees would benefit from the use of computers to create a visual environment with devices conveying touch sensation (haptics) to realistically mimic procedures on patients [3]. Removal of this initial experience from the clinical environment would be time efficient while improving patient safety and reducing the time taken for medical trainees to attain and maintain competence.

The reliability and feasibility of the performance of radiology resident trainees is assessed with two methods: direct observation of procedural skills and multisource feedback. The methods can yield reliable scores with appropriate sampling. The time required for direct observation of procedural skills is the duration of the procedure being assessed plus an additional one third of this time for feedback. In one study [4], the mean time required for each rater to complete the multisource feedback form was approximately 6 minutes. The methods are feasible and can be used to make reliable distinctions between physicians' performances. They also may be appropriate for assessing the workplace performance of other grades and spe-

Direct Observation of Procedural Skills

cialties of physicians, aiding formative assessment. Another study [5] confirmed that many medical students have not been directly observed in clinical training and that those observed more often expressed more self-reported confidence. Use of assessment measures that focus on direct observation and feedback during student-patient encounters may improve students' confidence.

Multisource feedback is the objective systematic collection of data and provision of feedback about an individual's performance from a number of raters from a variety of backgrounds (e.g., clinical colleagues, nurses, radiographers, and clerical staff) working with the individual. This method allows assessment of generic skills such as communication.

High reliability of an assessment process means that it would reach the same conclusion if it were possible to administer the same test again to the same individual in the same circumstances or at the least that the ranking of the best- to the worst-scoring students would not change. The assessment must be reproducible. Reliability is expressed as a coefficient varying between 0 (no reliability) to 1 (perfect reliability). In many assessments, the Cronbach alpha coefficient is used as an indicator of reliability. An appropriate cutoff for high-stakes assessments is usually 0.8. One factor for improving reliability is to increase the testing time to ensure wide content sampling and sufficient individual assessments by different assessors. It has been found that the reliability of multiple choice questions increases from 0.62 after 1 hour of testing to 0.93 after 4 hours and that the reliability of an oral examination increases from 0.50 after 1 hour to 0.82 after 4 hours.

Validity is a concept that cannot be measured but is an indicator of whether an assessment tests what it is meant to test. A number of forms of validity have been described, implying that multiple sources of evidence are required to evaluate the validity of an assessment. In addition to the reliability and validity of an assessment system, the educational impact, cost efficiency, acceptability, and feasibility should be evaluated. Optimizing an assessment method is about balancing these components. High-stakes pass or fail examinations need high reliability and validity, whereas a formative developmental assessment relying more on feedback to a trainee can focus more on impact and less on reliability.

Further studies are needed on the validity of direct observation of procedural skills.

Authors have commented on the lack of studies assessing the validity and reliability of direct observation of procedural skills, even though the method is fairly widely used to assess the competency of residents. In a 2003 review, Wilkinson et al. [6] found that no validated methods of procedural performance assessment were described in the literature. It is anticipated, however, that several studies of direct observation of procedural skills will be conducted as part of the introduction of the Foundation Programme in the United Kingdom. Thus evidence regarding instrument quality should soon emerge.

Despite the lack of evidence on its quality, direct observation of an individual's procedural skills certainly has high face validity. Examinees are observed in a situation that closely resembles clinical practice because the patients are real and the procedures are selected from normal routine. The only real authenticity issue is that physicians may not perform according to their usual standards because of the anxiety of knowing they are being assessed. The knowledge of being observed may influence behavior, so it can be argued that this method assesses not performance but maximum competence. Despite this criticism, the Royal College of Physicians, who developed a direct observation of procedural skills instrument for the Foundation Programme, anticipates that it will be found highly valid and reliable, particularly compared with the previous logbook system.

In addition to the studies on validity, further studies on the reliability of direct observation of procedural skills are needed. The main issues appear to be determining the number of procedures that should be observed to achieve adequate reliability and determining appropriate checklists and rating scales for different procedures. The number of encounters needed to ensure adequate reliability will be addressed in pilot studies for the Foundation Programme [7].

In terms of determining appropriate checklists and rating scales for direct observation of procedural skills, one issue is the degree to which these instruments should be structured. Studies in which the use of checklists and the use of global rating scales were compared in the context of standardized patient examinations have shown global ratings to be more reliable, suggesting that some degree of flexibility improves reliability. For the assessment of procedural skills, however, which are perhaps more mechanistic than other clinical skills, a structured approach may be required.

In a survey of a small group of anesthesiologists, Greaves and Grant [8] found that assessment of procedural skills, as opposed to more general clinical skills, is more effective when structured observations are made and the tasks are broken down into their components. Another variation of direct observation of procedural skills is the integrated procedural performance instrument. With that method, candidates are observed remotely with a video camera by assessors who collect data on the performance and send it to the supervising clinician. The main advantage of the integrated procedural performance instrument is that observing each performance remotely minimizes the effect that direct observation can have in altering the performance of a trainee owing to anxiety. Therefore, the assessment more closely approximates clinical reality. It is, however, highly resource intensive [9].

Methods

The resident chooses the timing, procedure, and assessor for the assessment and is responsible for ensuring completion of the required number of assessments. Direct observation of procedural skills is performed according to the steps developed for radiology. The steps in the procedure are observed by a faculty member and scored as feedback for formative assessment. The procedures are selected according to the time required to perform them and according to the level of training of the resident. Second-year residents may be required to perform drainage procedures, and residents in the third year and above may perform core biopsies. The time required for these procedures ranges from 20 to 30 minutes.

Applicable Procedures

In direct observation of procedural skills, residents may perform the following imaging-guided procedures: diagnostic aspiration of ascites and pleural effusion; catheter placement for ascites; therapeutic aspiration of pleural effusion; aspiration of liver abscess; catheter placement for postoperative collection; fine-needle aspiration of superficial organs (thyroid and lymph nodes) for cytologic evaluation; lumbar puncture under fluoroscopic guidance; liver biopsy; prostate biopsy; and breast biopsy.

Feedback is given for identification of agreed strengths and areas for development. The weak areas should be overcome with further dedicated learning, observation, and practice. The sole purpose of scoring

is provision of meaningful feedback. Areas of strength and those requiring further improvement are identified. Feedback must be given immediately after the assessment.

Training of assessors is important for proper evaluation and provision of feedback. The assessor must be familiar with the form developed for direct observation of procedural skills in radiology (Appendix 1) [10] and have expertise in the procedure being performed. The encounter should include a procedure in which a resident is normally expected to be proficient. Completion of the evaluation form should address the number of procedures the resident has completed before undergoing the assessment. The difficulty of procedure for the resident's level is scored. The assessor and resident are expected to read the guidance notes developed for the procedure. The scoring must be done with the full range of the rating scale. The rating scale has six steps because reliability decreases with the use of fewer categories [11]. A choice of not applicable is available if a step is not part of the procedure. The scale is bipolar and has a neutral point in the middle. Labeling of steps is for ease of marking the evaluation form.

Common Errors in Rating

There are common errors in ratings, and special precautions are needed to avoid these errors. The errors include personal bias and the halo effect [12].

Personal bias—Personal bias errors occur when the rater develops a tendency to rate all residents at approximately the same position on the scale. The problem arising from such a rating is that reliable discriminations are not made because scores are close to one another. In generosity error, the rater tends to rate all at the high end of the scale. In severity error, the rater prefers the lower end of the scale for everyone. In central tendency error, the rater avoids both extremes and rates everyone as average.

Halo effect—The halo effect is a general impression influence on rating. A rater who has a good or a bad impression of the resident tends to rate high or low regardless of actual performance. Concealing identity is not possible in a performance-based evaluation. Awareness of one's personal bias and prejudice can help avoid the halo effect. Errors can be reduced by proper design and use of the assessment instruments.

Ethical Issues

The ethical issues expected to be encountered in direct observation of procedural

skills in radiology include informed consent and proper explanation of the procedure. The assessment is intended to include only radiologic procedures that residents usually perform according to their level of training and education. There should never be a perception on the part of patients that they are experimental subjects. The resident and the assessor are expected to inform the patient that what is taking place is routine and that residents are performing procedures under supervision of faculty at all times. The only difference is that a formal assessment of the procedure is taking place through scoring of the performance. Patients always should have the choice not to participate.

Analysis

Reliability

Interrater reliability—To establish interrater reliability, two raters assess the same procedure by the same resident at the same time or an equivalent procedure performed after a time interval.

Stability and equivalence—Also known as test-retest reliability, stability and equivalence are established when a similar procedure is performed by the same resident after a time interval. The rater may be the same, and the procedure can be an equivalent one, such as aspiration of ascites and aspiration of pleural effusion.

Interitem reliability—Internal consistency refers to and is used to assess the degree of consistency among the items in a scale and among different observations used to derive a score. Most of the items selected in direct observation of procedural skills in radiology have a similar level of difficulty.

Validity

Face validity—Face validity simply means validity at face value. As a check of face validity, test items are sent to teachers to obtain suggestions for modification. Because of the vagueness and subjectivity of face validity, psychometricians abandoned this measure for a long time, but the concept has been resurrected in another form. Lacity and Jansen [13] define validity as making common sense, being persuasive, and seeming right to the reader.

Content validity—Content validity is used to assess whether a system measures the extent of knowledge it is intended to measure, that is, whether it contains the material that should be present for the training it is intended to impart. This type of validity is relatively easy to meet. In content validity, evidence

Direct Observation of Procedural Skills

is obtained by looking for agreement in judgments by judges. Face validity can be established by one person, but content validity should be determined by a panel.

Predictive validity—Predictive validity is used to determine whether performance on the assessment tool (e.g., direct observation of procedural skills in radiology) accurately and positively correlates with performance in practice (e.g., complex interventional radiology procedures). For ethical reasons, predictive validity is the most difficult level of validation to accomplish. Many systems never meet this level of validation, even though they are otherwise completely acceptable for training.

References

1. Lindsell D. Changes in postgraduate medical education and training in clinical radiology. *Biomed Imaging Interv J* 2008; 4:e19
2. Johnson S, Healey A, Evans J, Murphy M, Crawshaw M, Gould D. Physical and cognitive task analysis in interventional radiology. *Clin Radiol* 2006; 61:97–103
3. Collaborators in Radiological Interventional Virtual Environments. Interventional radiology (IR). CRAIVE Website. www.craive.org.uk/CRAIVE. Accessed March 3, 2010
4. Wilkinson JR, Crossley JG, Wragg A, Mills P, Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Med Educ* 2008; 42:364–373
5. Chen W, Liao SC, Tsai CH, Huang CC, Lin CC, Tsai CH. Clinical skills in final-year medical students: the relationship between self-reported confidence and direct observation by faculty or residents. *Ann Acad Med Singapore* 2008; 37:3–8
6. Wilkinson J, Benjamin A, Wade W. Assessing the performance of doctors in training. *BMJ* 2003; 327:S91–S92
7. Davies H, Archer J, Heard S. Assessment tools for Foundation Programmes: a practical guide. *BMJ Careers* 2005; 330:195–196
8. Greaves JD, Grant J. Watching anaesthetists work: using the professional judgement of consultants to assess the developing clinical competence of trainees. *Br J Anaesth* 2000; 84:432–433
9. Kneebone R, Nestel D, Yadollahi F, et al. Assessing procedural skills in context: exploring the feasibility of an Integrated Procedural Performance Instrument (IPPI). *Med Educ* 2006; 40:1105–1114
10. Royal College of Radiologists. Clinical radiology pilot: radiology direct observation of procedural skills (RAD-Dops). www.rcr.ac.uk/docs/radiology/pdf/radiologydopsfinalversion.pdf. Accessed March 3, 2010
11. Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. Oxford, UK: Oxford University Press, 2003:37–40
12. Linn RL, Miller MD. *Measurement and assessment in teaching*, 9th ed. Upper Saddle River, NJ: Pearson/Merrill Prentice Hall, 2005:267–269
13. Lacity M, Jansen MA. Understanding qualitative data: a framework of text analysis methods. *J Manage Inf Syst* 1994; 11:137–160

(Appendix appears on next page)

APPENDIX I: Evaluation Form for Direct Observation of Procedural Skills in Radiology

**Clinical Radiology Pilot
Radiology Direct Observation of Procedural Skills (RAD-Dops)**

Trainee's Code in pilot:

Assessor's Registration Number (e.g GMC, NMC):

Date of assessment (DD/MM/YY): / /

Year of specialty training: 1 2 3 4 5 6

Clinical setting: Ultrasound Computed Tomography Paediatric Imaging Magnetic Resonance Imaging
 Interventional Radiology Breast Imaging Fluoroscopy Other (please specify below):

Other setting:

Procedure name:

Number of times this procedure previously performed by trainee: 0 1-4 5-10 >10

Difficulty of procedure Low Medium High

	<i>Well below expectation for stage of training</i>	<i>Below expectation for stage of training</i>	<i>Borderline for stage of training</i>	<i>Meets expectation for stage of training</i>	<i>Above expectation for stage of training</i>	<i>Well above expectation for stage of training</i>	<i>Unable to comment</i>
1. Demonstrates understanding of indications, relevant anatomy and technique	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. Explains procedure/risks to patient, obtains/confirms informed consent where appropriate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. Uses appropriate analgesia or safe sedation/drugs	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. Usage of equipment	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. Aseptic technique	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. Technical ability	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. Seeks help if appropriate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8. Minimises use of ionising radiation for procedures involving x-rays	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9. Communication skills with patient/staff	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10. Quality of report of procedure	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Assessor's comments - if you have noticed anything especially good or needing further development please note it here:

Trainee's comments

