Medical Teacher, Vol. 26, No. 4, 2004, pp. 326–332

Taylor & Francis healthsciences

Measurement of perception and interpretation skills during radiology training: utility of the script concordance approach

LUCIE BRAZEAU-LAMONTAGNE[1], BERNARD CHARLIN[2], ROBERT GAGNON[2], LOUISE SAMSON[2] & CEES VAN DER VLEUTEN[3]

[1]University of Sherbrooke, Canada; [2]University of Montreal; Canada; [3]Maastricht University, The Netherlands

SUMMARY *Imaging specialties require both perceptual and interpretation skills. Except in very simple cases, data perception and interpretation vary among clinicians. This variability makes for difficulty in measuring these skills with traditional assessment tools. The script concordance approach is conceived to allow standardized assessment in contexts of uncertainty. In this exploratory study, the authors tested the usefulness of the approach for assessment of perceptual and interpretation skills in radiology. A perception test (PT) and an interpretation test (IT) were designed according to the approach. Both tests used plain chest X-rays. Three groups were tested: clerkship students (20), junior residents (R1–R3; 20), senior residents (R4–R5; 20). Eleven certified radiologists, all currently appointed to chest reading, provided the answers by aggregate scoring method. Statistics included descriptive, ANOVA, regression analysis, Pearson and Spearman correlation coefficients. Cronbach alpha values were 0.79 and 0.81 for the PT and IT respectively. Score progression was statistically significant in both tests. Perception scores progressed more rapidly than interpretation scores during training. Effect size was large in discriminating low versus higher level of expertise, 2.2 (PT) and 1.6 (IT). The Pearson correlation coefficient between both tests was 0.58. Cronbach alpha coefficient values indicate reasonable reliability for both tests. The linear progression of scores, each at its own pace, and the positive and moderate magnitude of the Pearson correlation coefficient are arguments suggesting measurement of two different skills. More studies are necessary to document the approach usefulness for assessment in radiology training.*

Introduction

Visual clinical specialties require both perceptual skills, which are mostly non-analytic, and interpretation skills, which look for clues and make a series of value judgments in order to arrive at a diagnosis (Norman *et al.*, 1992). Experience shows that residents' perceptual and interpretation skills do not necessarily develop synchronously. Knowing what to look for does not guarantee against 'creative reading', interpreting composite shadows for real nodules, for instance. Perception–interpretation discrepancies are common difficulties encountered in training residents in radiology. So far, such discrepancies remain resistant to objective demonstration and there is a need in radiology training programs for tests that can document the progress of students and residents in both skills.

One reason for the difficulty in achieving reliable tests of reading skills might be the variability that expert radiologists demonstrate when perceiving and interpreting diagnostic images. Research on clinical reasoning has demonstrated that, in similar clinical settings, physicians do not collect the exact same data and do not follow the same path of thought, even if they come to the same diagnosis (Grant & Marsden, 1988). Moreover, physicians perform with substantial variation on any specific real or simulated case (Barrows *et al.*, 1978; Elstein *et al.*, 1978). Among experts, unanimous reasoning on real clinical situations is the exception. Divergent opinion among them is rather the rule, even if they generally agree on the outcome, for instance the diagnosis. When translated into assessment settings, this implies that test answer grids cannot be (and most of the time are not) based on a single examiner (Swanson *et al.*, 1987). The script concordance approach (Charlin *et al.*, 2000a) offers a way to overcome these difficulties. It rests on three principles, each of them concerning one of the three components (Norman *et al.*, 1996) of all tests: the task required from examinees, the way examinees answers are recorded, and the way examinees' performance is transformed into a score.

The task presented to the candidates is challenging. It represents a real clinical situation usually described in a vignette (Charlin *et al.*, 2000a). Several options (diagnosis, management or attitude) are relevant to the situation and items are made with the questions experts ask themselves to progress toward a solution. For a test on interpretation in radiology, the task is based on an authentic set of images, presenting a genuine diagnostic challenge, even for an expert. Items ask how a specific sign, present (positive sign) or absent (negative sign), affects one of the hypotheses relevant to the situation. Items have three parts. The first presents the hypothesis. The second presents a sign (positive or negative) that may have an effect on the hypothesis. The third part, a Likert scale, captures examinees' answers. This response format is in accordance with what is known from clinical reasoning processes (Barrows *et al.*, 1978; Elstein *et al.*, 1978; Grant & Marsden, 1988). It allows for the measurement of the judgments that are constantly made within this process

Correspondence: Bernard Charlin, URDESS, Faculté de Médecine-Direction, Université de Montréal, C.P. 6128, succursale centre-ville, Montréal, Québec, H3C 3J7 Canada. Tel: 514 343 7827; fax 514 343 7650; email: bernard.charlin@umontreal.ca

ISSN 0142–159X print/ISSN 1466–187X online/03/030326-7 © 2004 Taylor & Francis Ltd
DOI: 10.1080/01421590410001679000

Interpretation test: Nodule

On this X-ray, a pulmonary nodule is seen in the anterior segment at the left upper lobe.
The following diagnoses are included.

Lung cancer

Histoplasmosis

Solitary metastasis

Rheumatoïd nodule

Pulmonary abscess

On the X-ray, the following signs are also seen:

1. No calcification is seen in the nodule. What effect this finding has on the diagnostic possibilities?

Primary lung cancer	–3	–2	–1	0	+1	+2	+3
Histoplasmosis	–3	–2	–1	0	+1	+2	+3
Solitary metastasis	–3	–2	–1	0	+1	+2	+3

2. The nodule is not cavitary

Solitary metastasis	–3	–2	–1	0	+1	+2	+3
Rheumatoid nodule	–3	–2	–1	0	+1	+2	+3
Pulmonary abscess	–3	–2	–1	0	+1	+2	+3

Answers grid (circle the right answer)

–3 The diagnosis is excluded	+1 The diagnosis is a bit more probable
–2 The diagnosis is a lot less probable	+2 The diagnosis is a lot more probable
–1 The diagnosis is a bit less probable	+3 The diagnosis is the only one possible
0 No effect on hypothesis	

Figure 1. Examples of items concerning the first set of X-rays in the interpretation test.

(Charlin *et al.*, 2000b). An illustration of the format is given in Figure 1. The method of building tools according to the script concordance approach is described in detail elsewhere (Charlin *et al.*, 2000a).

The scoring method takes into account variation of answers among jury members. It is an adaptation of the aggregate scoring method (Norman, 1985; Norcini *et al.*, 1990). Credits on each item are derived from the answers given by a panel of reference. The principle is that any answer of an expert reflects a valid opinion that should be taken into account, even those with poor agreement among experts. The credit for each answer is the number of panel members that have provided that answer, divided by the modal value for the item. For example (see Figure 2), if on an item six panel members (out of 11) have chosen response −1, this choice receives 1 point (6/6). If three experts have chosen response −2, this choice receives 0.5 (3/6), and if two experts have chosen response +1, this choice receives 0.33 (2/6). The total score for the test is the sum of credits obtained on all items. This score is then divided by the number of items and multiplied by 100 to get a percentage score.

The script concordance approach uses a comparison of examinees' answers with those of a reference panel of experts (11 experts in our study). It assesses whether examinees' knowledge organization for clinical tasks, their script (Charlin *et al.*, 2000b), is in concordance with the scripts of the reference panel of experts. It allows probing of the clinical reasoning process, instead of focusing on the diagnostic outcome alone, as traditional examination formats (such as multiple-choice question test) would do.

The use of the script concordance approach has previously been tested with scenarios of radiological problems presented in written vignettes (Charlin *et al.*, 1998). Results show that it is possible with these written descriptions to detect interpretation skill progression with training in radiology. The present study was undertaken to verify whether this effect could be found with the use of actual X-rays, instead of

	–3	–2	–1	0	+1	+2	+3
No. of experts' answer	0	3	6	0	2	0	0
Raw score	0	3/6	6/6	0	2/6	0	0
Student credit for the item	0	0.5	1	0	0.33	0	0

Figure 2. Method of score transformation.

Perception test: Nodule

Is there:	Yes	No	Artefact	Don't know
A nodule in the right upper lobe?	☐	☐	☐	☐
A nodule in the left upper lobe?	☐	☐	☐	☐
A mass in the aortic pulmonary window?	☐	☐	☐	☐

Figure 3. Example of items from the first set of X-rays in the perception test.

written scenarios. We also wanted to disentangle perception from interpretation skills and to test the application of the script concordance approach to visual perception of radiological signs. The participant had to decide whether a sign was present or not on the films (see Figure 3). Answer options were yes (the sign is present), no (the sign is not present), artifact (it is an artifact) or I don't know. A pilot study showed that experts' opinions on signs present or absent on films were far from unanimous, so the use of the aggregate method for scoring was appropriate.

Research questions were as follows: (1) Does the script concordance approach allow reliable and valid measurement of perception and interpretation skills in radiology settings? (2) How do scores progress along training on the two skills?

Method

Subjects

The study was carried out in two different teaching radiology departments. Trainees of three different training levels accepted to participate freely: clerkship students (20), junior residents (R1 to R3; 20), and senior residents (R4 and R5; 20). To be selected, clerkship students had to have completed an undergraduate rotation in general radiology. Recruitment stopped when 20 subjects in each category were enrolled. Eleven staff members from university departments—all currently appointed to chest reading—agreed to be the panel of reference.

Material

Both tests were drawn from the plain chest X-ray domain: Four common presentations were selected, each one referring to a different training objective shared by both undergraduate and postgraduate radiology programs: coin lesion, atelectasis, interstitial infiltrate, and mediastinal mass. Four sets of PA and lateral films were used in the interpretation test and four others were used for the perception test. All were representative exemplars of each diagnostic challenge.

Interpretation test (IT)

Since the IT focuses on sign interpretation and not on sign perception, the actual signs were clearly specified on test sheets. The introducing paragraph was the following: 'On this film, there is a nodule in the anterior segment of the left upper lobe. Hypotheses are: primary neoplasm, histoplamosis, metastatic nodule, necrobiotic nodule, and pulmonary abscess. There are other signs on the films. What effect does each of them have on the diagnostic hypotheses you consider?' Positive (e.g. there are calcifications in the nodule), or negative (e.g. the nodule is not cavitated) signs were presented. The task was to decide what effect each sign had on the current hypothesis (see Figure 1). The answer format was a seven-point Likert scale ranging from −3 'the diagnosis is excluded' to +3 'the diagnosis is the only one possible', with 0 corresponding to 'no effect on hypothesis'. The whole IT was made of 145 items: case A (coin lesion, 45 items), case B (atelectasis, 35 items), case C (interstitial infiltrate, 35 items), case D (and mediastinal mass, 30 items).

Perception test (PT)

The PT was based on the same four problems (coin lesion, atelectasis, interstitial infiltrate, and mediastinal mass), each one pictured on a set of chest X-rays (different sets from those used in the IT). The PT was made of 38 items. For each case a series of radiological signs was provided. The participant had to decide if the sign was present or not on the films (see Figure 3). Answer options were yes (the sign is present), no (the sign is not present), artifact (it is an artifact) or I don't know. Scores on both tests were computed from answers given by a reference panel of 11 experts. Panel members were asked to complete the tests individually.

Statistical analysis

Analyses on the IT were done at the level of signs to prevent artificial inflation of reliability coefficients due to item dependence related to a single sign (there are five items for each sign). Thus the mean score of the five items related to each sign was considered as the unit of measurement. Data were analyzed at two levels: the problems (four cases) and the whole test (summation of scores of the four problems). All global scores (cases and total) were plotted to a common denominator of 100 points to be easily comparable.

Statistical analyses included descriptive statistics, ANOVA, and linear regression analysis. Scale reliability was evaluated by alpha coefficient for internal consistency (standard Cronbach alpha) without any form of optimization. One-way analysis of variance was used to test the differences between the three groups on global score and cases. Simple regression analysis was used to compare 'progression' of scores with increased level of training (used as the independent variable) and to estimate the slope of the regression line (unstandardized regression coefficient). A Z test (Kanji, 1993) was used to compare the two regression coefficients.

The relationship between interpretation and perception tests was assessed with the Pearson correlation coefficient, while the relationship between level of expertise and performance was assessed with Spearman's coefficient. Effect size was calculated using differences between mean of extreme groups (clerkship students versus senior residents) and standard deviation of the lowest mean group. All tests were two-sided with an alpha level of 0.05.

Results

Data were obtained from three groups of 20 trainees. Results are presented at the level of the whole test, then at the level of the cases because of the well-known phenomenon of content specificity in clinical reasoning (Norman *et al.*, 1985).

Analysis at the whole test

The alpha coefficient on the 38 items of the perception test was 0.79. It was of the same order (0.81) for the 145 items regrouped at the level of 29 signs in the interpretation test. As expected, there were high item–total correlations for both tests. The correlation between both tests was 0.58 ($p < 0.05$). For both tests, the difference between the three groups is statistically significant (<0.001) for the test as a whole. Effect size is very large for both tests, 2.2 (perception test) and 1.6 (interpretation test).

The slope of the relationship between level of training and the mean scores on the whole test was estimated using the unstandardized regression coefficients. The slope is steeper (has higher value) in the perception test (slope = 10.9) than in the interpretation test (slope = 5.7). Figure 4 illustrates the progression of global scores for both tests.

Analysis at the level of cases

When using the four cases as the unit of analysis, the alpha coefficients are 0.77 for the perception test and 0.72 for the interpretation test.

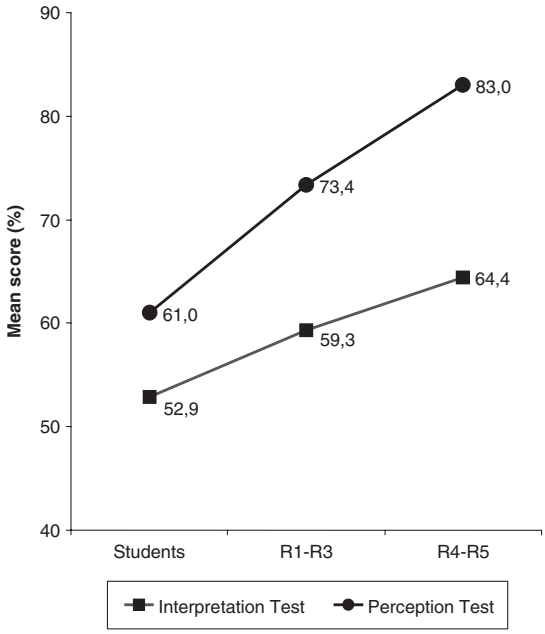

Figure 4. Evolution of perception and interpretation scores.

Table 1. Correlation matrix for cases and total scores on both tests[a].

	Perception					Interpretation			
	Case A	Case B	Case C	Case D	Total	Case A	Case B	Case C	Case D
Perception									
Case A	–								
Case B	0.28	–							
Case C	0.47	0.59	–						
Case D	0.41	0.58	0.41	–					
Total	0.63	0.83	0.82	0.78	–				
Interpretation									
Case A	0.44	0.61	0.60	0.35	0.65	–			
Case B	0.25	0.42	0.31	0.21	0.39	0.64	–		
Case C	0.20	0.29	0.23	0.17	0.29	0.39	0.56	–	
Case D	0.14	0.29	0.26	0.11	0.27	0.26	0.30	0.26	–
Total	0.38	0.57	0.51	0.30	0.58	0.85	0.85	0.91	0.55

Note: [a]Correlations higher than 0.25 are statistically significant $p < 0.05$.

Table 2. Performance of groups on the perception test.

	Case A Mean (SD)	Case B Mean (SD)	Case C Mean (SD)	Case D Mean (SD)	Total test Mean (SD)
Students	66.9 (13.5)	65.7 (15.1)	59.6 (16.9)	53.0 (15.7)	61.0 (10.2)
R1–R3	74.5 (12.7)	81.6 (14.7)	74.1 (13.4)	63.8 (12.8)	73.4 (8.8)
R4–R5	82.1 (8.0)	93.6 (6.1)	85.4 (7.2)	70.8 (9.6)	83.0 (4.0)
Anova F	8.5	24.3	19.3	9.6	36.8
p	<0.001	<0.001	<0.001	<0.001	<0.001

Table 3. Performance of groups on the interpretation test.

	Case A Mean (SD)	Case B Mean (SD)	Case C Mean (SD)	Case D Mean (SD)	Total test Mean (SD)
Students	56.5 (8.8)	45.7 (11.8)	53.3 (7.9)	55.5 (10.3)	52.9 (7.1)
R1–R3	69.5 (11.6)	50.4 (9.9)	54.6 (9.9)	59.7 (11.4)	59.3 (7.9)
R4–R5	76.2 (8.2)	57.6 (10.9)	58.8 (7.9)	60.9 (10.4)	64.4 (6.2)
Anova F	21.6	6.1	2.3	1.4	13.0
p	<0.001	0.004	0.11	0.26	<0.001

Table 1 shows the correlations between scores on cases and total scores for both tests. Moderate correlations were found between cases within each test (0.28 to 0.64) and low to moderate correlations between cases across tests (0.11 to 0.61). Moderate correlations are observed between cases from both tests (0.30 to 0.57). Case A of the interpretation test has a very special profile since it shared high correlation (0.60 and 0.61) with case B and case C of the perception test and with global perception score (0.65). Correlation coefficients between tests results show moderate relationships for case A (0.44; $p < 0.01$); case B (0.42; $p < 0.01$); case C (0.21; $p = 0.08$); case D (0.11; $p = 0.42$) and for global score (0.58; $p < 0.01$).

Mean scores obtained for each of the four problems by the three groups with the perception test and the interpretation test are given in Table 2 and Table 3. In the perception test, all differences between groups are statistically significant (<0.001) for each of the four problems. In the case of the interpretation test, significant differences between the three groups are found for the two first problems. Smaller (non-significant) differences between groups were noted for problem 3 ($F = 2.3$; $p = 0.112$) and for problem 4 ($F = 1.4$; $p = 0.215$). In both tests, with all four problems, all the means showed a constant progression from the student training level to the senior resident level (R4–R5).

In general, the slope is steeper in the perception test (case A = 7.6; case B = 13.9; case C = 12.9; case D = 8.9) than in the interpretation test (case A = 9.8; case B = 5.9; case C = 2.8; case D = 2.7). Case A is the exception, as the slope is steeper in the interpretation test. On the other hand, regression coefficients for each case of the perception test and of the interpretation test are not statistically significant (all p values >0.10).

Estimation of Spearman correlation shows strong relationship between level of expertise and performance on perception test: case A ($r = 0.34$); case B (0.77); case C (0.66); case D (0.63), total score (0.79). The relationship is lower for the interpretation test: case A (0.64); case B (0.41); case C (0.25); case D (0.24) and total score (0.57). Correlations higher than 0.25 are statistically significant at an alpha level of 0.05.

Discussion

Problems that professionals handle in their practice are of two kinds: technical and professional (Schön, 1983). Technical ones are those resolved by applying known 'algorithmic' strategies, in defined and stable contexts, with clear and predictable goals. They are called well-defined problems. In medicine, each patient is embedded in a unique context, depicted with often imperfect, inconsistent, or even inaccurate information. Translated to the radiology context, images are far from being all being 'standard'. Hence, radiologists, like other clinicians, are often challenged with the second or ill-defined type of problem (Harris, 1993). According to Schön (1983) the knowledge required to solve ill-defined problems differs from the technical knowledge used to handle well-defined problems. He called it professional knowledge. He stated that professional schools emphasize learning and assessing technical knowledge too much while often neglecting the crux of the matter, i.e. professional knowledge.

When reasoning on complex problems, professionals follow different paths on their way to the solution, as shown repeatedly in research on clinical reasoning (Barrows *et al.*, 1978; Elstein *et al.*, 1978; Grant & Marsden, 1988). It has also been shown that perception and interpretation of features in visual domains is context sensitive (Brooks *et al.*, 2000). Nevertheless, assessment jury members are still commonly required to provide 'the correct' answers examinees should arrive at for each item tested. Therefore examinees are submitted to problems defined well enough to allow clear-cut 'correct' answers, and are almost never exposed to the type of ill-defined problems clinical experts deal with in daily practice. The script concordance approach,

because it gives credit to divergent ways of thinking, overcomes this limitation of traditional exams.

Both tests used in the study demonstrate good metric qualities. The interpretation test is reliable with an acceptable value of alpha coefficient (0.81), even if it was computed at the level of signs (29), and not at the level of items (145), to prevent an artificial inflation of the coefficient due to high inter-item correlations among same case items. The test is well accepted by examinees, who liked its 'real clinical life' style. It is composed of authentic questions that experienced clinicians ask themselves while solving a problem. It has also the advantage of a standardized scoring process. The other test (perception test) also shows good reliability (0.79) with few items (38). It discriminates examinees efficiently as shown by high F values, large effect size and steep slopes for regression lines. Both instruments carry high discriminative power as effect size is large in both tests, 2.2 (perception test) and 1.6 (interpretation test). Therefore it is possible to detect variations between small groups of examinees.

Perception and interpretation skill scores are good predictors of levels of training. Performance on the perception test stands as the best predictor. Patterns observed in the correlation matrix show that both measures are substantially correlated and that competence in the perception realm emerges as the best predictor of level of training. Both competences increase linearly at their own progression pace. Study data seem to indicate that perception skill is acquired sooner than interpretation skills. At the beginning of training, it seems easier to perceive than to interpret.

This exploration study has several limitations. (1) The transverse nature of the present study limits the possibility to document the true progression of both skills. (2) Small sample size limits the possibly to detect moderate differences between the slope of the two progression lines. (3) While data provide arguments on measurement of two different skills, one has to establish more clearly that the tests are really measuring perception and interpretation skills. For instance, asking 'is there a nodule present' may elicit different responses from a free-form question, such as 'is there an abnormality'. A study testing the equivalence of responses between the two types of question will have to be undertaken. At this time the gold standard in evaluation in radiology training is oral where examinees are asked by an observer to read films and are probed on perception and interpretation of radiological signs. A study is needed to demonstrate that PT and IT scores correlate strongly with observers' assessment on the skills of examinees.

Conclusion

This study indicates that it may be possible, with the script concordance approach, to reliably and validly measure perception and interpretation skill. More studies are necessary to document its usefulness for assessment in radiology training. If the validity of these tools is confirmed, their use in radiology training may allow, by using real-life stimuli and tracking answers with standardized methods, follow up of the acquisition of both skills along training. Low scores on the interpretation test might help identify those students or residents who would benefit from special educational assistance to improve their problem-solving skills, while the use of a perception test would permit identification of those residents in need of perceptual remedies.

Practice points

- The script concordance approach makes standardized assessment possible in contexts of uncertainty.
- Examinees' answers are compared with those of a panel of experts, with a method that takes in account variability of experts' answers.
- Study results indicate a way to make reliable and valid measurement of perception and interpretation skills in radiology training.
- If these results are confirmed by subsequent studies, the use of perception and interpretation tests in radiology training may allow the following up of the acquisition of those skills during training and proposing, when needed, educational remediation activities.

Acknowledgments

This research project was funded by a grant from the Medical Research Council of Canada/Association of Canadian Medical Colleges.

Notes on contributors

LUCIE BRAZEAU-LAMONTAGNE, MD MA (Phil), is Professor of Radiology at University of Sherbrooke, Canada.

BERNARD CHARLIN, MD PhD, is Professor of Surgery and Director of the Unit of Research and Development in Education at the Faculty of Medicine of University of Montreal. His field of research is standardized assessment of ill-defined problems.

ROBERT GAGNON, MSc (Psy), is methodologist. He works as consultant and has a personal interest in assessment and cognitive psychology.

LOUISE SAMSON, MD, is Professor of Radiology at University of Montreal. She is Vice-President, Professional Development, of the Royal College of Physician and Surgeons of Canada.

CEES VAN DER VLEUTEN, PhD, is Professor and Chair of the Department of Educational Development and Research at the University of Maastricht, The Netherlands.

References

BARROWS, H.S., FEIGHTNER, J.W., NEUFELD, V.R. & NORMAN, G.R. (1978) Analysis of the Clinical Methods of Medical Students and Physicians. Final Report to the Province of Ontario Department of Health.

BROOKS, L.R., LEBLANC, V.R. & NORMAN, G.R. (2000) On the difficulty of noticing obvious features in patient appearance, *Psychological Science*, 11, pp. 112–117.

CHARLIN, B., ROY, L., BRAILOVSKY, C.A. & VAN DER VLEUTEN, C.P.M. (2000a) The Script Concordance Test: a tool to assess the reflective clinician, *Teaching and Learning in Medical Education*, 12, pp. 189–195.

CHARLIN, B., TARDIF, J. & BOSHUIZEN, H.P.A. (2000b) Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research, *Academic Medicine*, 75, pp. 182–190.

CHARLIN, B., BRAILOVSKY, C.A., BRAZEAU-LAMONTAGNE, L., SAMSON, L. & LEDUC, C. (1998) Script questionnaires: their use for assessment of diagnostic knowledge in radiology, *Medical Teacher*, 20, pp. 567–571.

ELSTEIN, A.S., SHULMAN, L.S. & SPRAFKA, S.A. (1978) *Medical Problem Solving: An Analysis of Clinical Reasoning* (Cambridge, MA, Harvard University Press).

GRANT, J. & MARSDEN, P. (1988) Primary knowledge, medical education and consultant expertise, *Medical Education*, 22, pp. 173–179.

HARRIS, I. (1993) New expectations for professional competence, in: L. Curry & J.F. Wegin (Eds) *Educating Professionals. Responding to New Expectations for Competence and Accountability*, pp. 17–52 (San Francisco, Jossey-Bass).

KANJI, G.K. (1993) *100 Statistical Tests*, p. 27 (Thousand Oaks, CA, Sage Publications).

NORCINI, J.J., SHEA, J.A. & DAY, S.C. (1990) The use of the aggregate scoring for a recertification examination, *Evaluation and the Health Professions*, 13, pp. 241–251.

NORMAN, G.R. (1985) Objective measurement of clinical performance, *Medical Education*, 19, pp. 43–47.

NORMAN, G., SWANSON, D.B. & CASE, S.M. (1996) Conceptual and methodological issues in studies comparing assessment formats, *Teaching and Learning in Medicine*, 8, pp. 208–216.

NORMAN, G., COBLENTZ, C., BROOKS, L. & BADCOOK, C. (1992) Expertise in visual diagnosis: a review of the literature, *Academic Medicine*, 67, pp. S78–S83.

NORMAN, G.R., TUGWELL, P., FEIGHTNER, J.W., MUZZIN, L.J. & JACOBY, L.L. (1985). Knowledge and clinical problem solving, *Medical Education*, 19, pp. 344–536.

SCHÖN, D.A. (1983) *The Reflective Practitioner: How Professionals Think in Action* (New York, Basic Books).

SWANSON, D.B., NORCINI, J.J. & GROSSO, L.J. (1987) Assessment of clinical competence: written and computer-based simulations, *Assessment and Evaluation in Higher Education*, 12, pp. 220–246.