

Chapter 39

Assessing clinical reasoning

C.P.M. van der Vleuten, G.R. Norman and L.W.T. Schuwirth

CHAPTER CONTENTS

Historical perspective	414
New concepts of clinical reasoning	414
New assessment developments	415
Key feature	416
Multiple choice questions (MCQs)	417
Other clinical reasoning formats	418
Other clinical reasoning assessment methods	418
Implications and advice for the teacher	419

The term *clinical reasoning* is used in varying ways. In this chapter we use it to refer to the mental activities involved in arriving at a diagnosis and a management plan. Thus it is related to activities like history taking or physical examination, which are somewhat distinct.

Typical for the assessment of clinical reasoning is the use of an authentic professional situation as a stimulus format, usually in the form of a simulation representing a professional situation using a paper, a verbal or a practical performance situation. Many representations are possible, ranging from simple to complex, and they can be connected to many different types of response format. Experimentation with all these phenotypes actually reflects the history of clinical reasoning assessment as described in more detail below. It follows very intuitive notions of how clinical reasoning should be assessed, moving towards increasingly more simplified forms of assessment based on growing insights from research and practical experiences. To a large extent, the history of clinical reasoning represents a sobering experience, falsifying many of the original intuitive beliefs. However, this is not uncommon in education research (van der Vleuten et al 2000) and really makes the story worth telling. In doing so, we will limit our discussion primarily to cognitive assessment methods. We acknowledge that clinical reasoning also occurs in performance-based measures such as the Objective Structured Clinical Examination (OSCE) (Petrusa 2002) or methods involving real-life clinical settings (Turnbull & Van Barneveld 2002); but reasoning is first and foremost an activity of the mind.

HISTORICAL PERSPECTIVE

In the 1960s and 1970s there was considerable interest in the development of methods which assessed 'clinical problem-solving skills'. The main thrust was to mimic authentic clinical situations as creatively as possible both in the stimulus and in the response format. This entailed a simulation on paper, and later by computer, of the process by which a doctor took a history, obtained information from the physical examination and made diagnostic, investigational and management decisions.

Undoubtedly the most popular of the many variants was the *Patient Management Problem* (PMP) (Mcguire & Babbott 1967). A typical PMP begins with a variable amount of information about the patient. The student is then requested to collect further data sequentially in either a linear or a branching fashion, typically using a 'rubout' pen that exposes the answer. After collecting history and examination data, ostensibly in the manner and order that would have pertained in the live patient situation, the student may select investigations and/or make diagnostic and management decisions. The pathway of the student is compared to that of an expert or criterion group, and composite scores are determined.

The death knell of PMPs was the finding that performance on one PMP is a poor predictor of performance on another PMP. From a number of studies the correlation across problems was of the order of 0.1–0.3 (Norman et al 1985). This observation appears to undermine one of the original hypotheses underlying the development of problem-solving simulations, i.e. that they measure problem-solving ability. If that were so, correlations between PMPs ought to be high, since those who are better problem solvers should exhibit superior performance across a wide range of problems, independent of specific content knowledge. The explanation of this phenomenon is referred to variously as *content specificity* or *case specificity* (Elstein et al 1978). Interestingly, the finding is not peculiar to PMPs but is also seen for other methods which assess aspects of clinical competence and performance (van der Vleuten & Schuwirth 2005).

One variant that has survived the passage of time is the modified essay question (MEQ), which

has been used quite extensively by the medical profession in some parts of the world, both for in-course assessments and for the certification of competence. This reflects, in part, the relative ease of construction of MEQs as compared to PMPs (Feletti & Engel 1980). A typical MEQ once again begins with a case vignette as a stimulus. Students are asked to respond to questions in a short essay format. New information is provided sequentially which relates to differing and evolving circumstances of the same case. Some skill is required to avoid providing cues to earlier or subsequent sections of the MEQ. Few studies are available of the reliability and validity of this method but it has face validity, appears to be acceptable and is practicable (Feletti & Engel 1980, Neufeld & Norman 1985). Nevertheless, there is no reason to presume that the MEQ would be any less vulnerable to the deleterious effect of content specificity than any other format.

Given these limitations, doubt has been cast on the value of any format which involves extensive and lengthy testing with relatively few cases (Swanson et al 1987). In addition, the experience with PMPs has alerted us to our limited understanding of the nature of clinical reasoning. Among other things, it has stimulated research of a more fundamental nature into the cognitive functioning of medical students and doctors (Eva 2005, Norman 2005).

NEW CONCEPTS OF CLINICAL REASONING

In the 1970s and 1980s several studies showed that while expert clinicians systematically outperformed less experienced doctors on a variety of simulations of clinical problem solving (Neufeld et al 1981), there was little difference in the problem-solving process they used. This led to a new direction in fundamental research, guided primarily by methods of cognitive psychology (Eva 2005, Norman 2005, Norman et al 1989, Regehr & Norman 1996, Schmidt et al 1990) (see also Chapters 10 and 20 in this book).

Current understanding would suggest that problem-solving ability is not a separate skill or

entity which grows with training and experience, and that it cannot be measured independently of relevant content knowledge. Problem-solving ability appears to be highly dependent on knowledge, not just the amount of knowledge but also its specificity and the way it is structured, stored, accessed and retrieved. This is not to say that knowledge alone is sufficient for efficient and effective clinical reasoning. Higher-order control processes also play an integral role (Bransford et al 1986). But the notion that there is a general, content-independent skill that experts acquire during training is simply incompatible with the evidence. Knowledge – its amount, its kind, and its organization – is central to expertise.

One theory of knowledge organization proposes three different kinds of knowledge relevant to solving clinical problems. The most elementary is knowledge of disease processes and causal relationships, the basic science of medicine. At a later level, students acquire *illness scripts* which are quite literal list-like structures relating signs and symptoms to disease prototypes (Feltovich & Barrows 1984). At the highest level of functioning, the expert uses a sophisticated form of pattern recognition characterized by speed and efficient use of information (Brooks et al 1991, Schmidt et al 1990). It appears that this representation is drawn to a large degree from direct experience with patients, and that pattern recognition is, in fact, recognition at a holistic level of the similarity between the present patient and previous patients (Hatala & Norman 1999).

This is not to indicate that all expert clinical reasoning occurs by pattern recognition. More recent research suggests, not surprisingly, that experts may make use of all kinds of knowledge – basic science, clinical and experiential (De Bruin et al 2005, Norman 2005). If the problem is one with which the person has had considerable previous experience, then it is probably recognized very early by a pattern recognition process. Little active thinking is required and there is a rapid resolution of the problem. However, if no easy solution is evident, more systematic intellectual activities must be brought into play, either formal testing of hypotheses through accumulation and weighting of specific data, or causal reasoning at the level of basic disease mechanisms. An

individual will demonstrate a range of approaches, both within and across problems, depending on previous experience and exposure to problems of a similar nature.

To the extent that this view is correct, it is evident that early attempts to assess clinical reasoning were doomed. We cannot consider it a generic process. Instead, we must contemplate the evaluation of several qualitatively different strategies. Some, like pattern recognition, are efficient and indeed may be over in seconds. These strategies will defy any attempt at measurement of the process. Some, like causal reasoning, are focused on detailed reasoning about mechanisms and are little concerned with data acquisition. As a result, they are inadequately captured by a focus on observable behaviours like history taking and physical examination. These issues have serious implications for assessment.

NEW ASSESSMENT DEVELOPMENTS

From the above experiences and empirical findings several things became clear. The first is that assessment must be anchored in case-based material presented in a way that will induce and sample clinical-reasoning activities. The second is that laboriously taking a student through the full data-gathering and investigational phase of a real or simulated clinical case is an inefficient approach when the concern is to evaluate clinical reasoning, simply because of the content-specificity problem and the consequent need to present students with large numbers of cases before satisfactory levels of test reliability can be achieved. For example, it has been shown because of this problem that up to 8 hours of testing time may be required to achieve reliable assessments with PMPs (Norcini et al 1985). Such studies have triggered a search for more cost-effective methods with simpler simulation technologies. We will discuss a number of them. There is one other implication; since there are multiple knowledge representations, each or all of which may be invoked to solve a particular problem, it makes little sense to attempt to identify the specific knowledge or strategy used to solve any problem. It suffices simply to ensure that sufficient numbers of cases have been

sampled to differentiate reliably between better and poorer clinical reasoners on the basis of their success rates.

In examining the various contemporary methods, one useful distinction in assessment methods is between stimulus formats and response formats (Norman et al 1996). The *stimulus format* refers to the task that is being presented to a candidate in the assessment. It may be very simple and short, for example a question about the signs and symptoms of a particular disease, or it may be very complex and time consuming. A case scenario or maybe even a video presenting a patient case to the candidate represents an illustration of the latter. The stimulus format is ended with a lead-in question that connects the previous information to required response from the candidate, for example, 'What is the most likely diagnosis?' The *response format* refers to the way the response of the candidate is captured. It could consist of a short menu of options (multiple choice), long extensive (computerized) menus, a short write-in format, a long write-in format (essay-type questions), an oral response (oral examinations) or a behavioural response either in a simulated environment (e.g. OSCE) or in a real-life context.

KEY FEATURE

As a suggestion from a 'think-tank' conference on clinical reasoning, the first Cambridge Conference, the idea emerged to focus on essential elements of a clinical case (Norman et al 1992). The idea was based on the premise that any single case contained much 'dead wood' from a clinical-reasoning perspective. For example, in one case the critical challenge might be to elicit and interpret elements within the history, with little further being added by the physical examination and laboratory investigations. In another case the challenge might be the appropriate selection and interpretation of laboratory results. In other words, it may be possible to focus the problem-solving stimulus. One concrete outcome has been the *key feature* approach developed for the Medical Council of Canada certification examinations as an alternative to PMPs (Page & Bordage 1995, Page et al 1990). In this procedure, clinical situations, as presenting in actual practice, are

produced as written case scenarios representing the stimulus format. The key features are identified on the basis of those elements critical to resolution of the problem. Questions relating to the key features are then devised and may be posed in a variety of response formats (e.g. short answer, multiple choice questions (MCQ) or selection from longer menus of options). Such an approach allows a sample of 40–50 cases to be administered in the same time as that required to administer 12–15 PMPs.

Studies so far have indicated improved reliability as compared to the PMP, but still 3–5 hours of testing time is required. Data from the Medical Council of Canada showed that a reliability of 0.80 is reached with approximately 40 cases in 4 hours of testing time (Page & Bordage 1995). Other studies reported slightly worse findings (Hatala & Norman 2002), or slightly better findings (Fischer et al 2005). A recent study has shown that 2–3 items per case is the optimal for achieving maximum reliability (Norman et al 2006); reading time will compromise reliability when fewer items are used and information redundancy will compromise reliability when more items are used. Validity studies investigating correlations with other measures typically show moderate correlations. More compelling are studies that use think-aloud strategies when comparing stimulus formats. They show that case-based stimulus formats elicit other cognitive processes than fact-oriented stimulus formats (Schuwirth et al 2001, Skakun et al 1994). Response formats that use menus instead of write-ins may cue the candidate to both correct and incorrect answers (Schuwirth et al 1996a) with slightly higher scores as a net effect, naturally depending on the number of alternatives in the menu (Schuwirth et al 1995). Score correlations across these response formats, however, are invariantly high (Schuwirth et al 1996a).

A modern variation of the key feature format is the use of computers for test administration, allowing more flexible use of pictorial and audio information (Schuwirth et al 1996b, Fischer et al 2005). Practical information on the construction of key features is readily available (Schuwirth et al 1999, Farmer & Page 2005). The writing of key features requires significant staff input (Hatala & Norman 2002).

MULTIPLE CHOICE QUESTIONS (MCQs)

In their simplest format simulations take the form of *vignette-based MCQs* (Case & Swanson 2002). This is the preferred format of the US National Board of Medical Examiners in their undergraduate licensure examinations. In recent years they completely changed the assessment strategy of their written examinations. All test items used are now vignette-based MCQs. The United States Medical Licensing Examination (USMLE) consists of two parts. Step 2 is the clinical component and is fully patient-based. Short cases are presented that require some form of judgement or decision. This may be related to data gathering, to case management or to any other phase of the clinical problem. For example, instead of asking:

Ibuprofen belongs to a certain group of NSAIDs. Which group?

- a. Salicylates
- b. Acetic acid derivatives
- c. Oxycam derivatives
- d. Propionic acid derivatives
- e. Pyrazolinone derivatives

this topic of pain management could be addressed as for example:

Mr Brown has a carcinoma of the esophagus. The carcinoma has metastasized and curative treatment is not possible. Initially, the disorder caused little pain, which was easily suppressed with nonsteroidal anti-inflammatory analgesics and a weak opioid. Due to more invasive growth of the carcinoma, the pain has increased and the pain management is no longer adequate even at the highest dosage of the current medication. Which is the most indicated next step in the pain therapy in this case?

- a. Adding a tricyclic antidepressant to the present medication.
- b. Adding a strong opioid to the therapy while discontinuing the weak opioid medication
- c. Increasing the dosage of the nonsteroidal anti-inflammatory analgesics
- d. Adding a tranquillizer to the current medication.

After the case presentation the lead-in prompts the candidate to make a choice from a menu.

USMLE Step 1 is on basic sciences, but even there the strategy is to design a reasoning question. Instead of asking:

Which neurotransmitter/s activate/s the sweat glands?

- a. Only acetylcholine
- b. Only adrenaline and noradrenaline
- c. Only adrenaline and acetylcholine
- d. Only noradrenaline and acetylcholine
- e. Noradrenaline, adrenaline and acetylcholine.

the topic of temperature control could be addressed as for example:

Charles and Irene are going to travel through Mexico for 2 months. At Mexico City airport the temperature is no less than 40°C. Their clothes get sticky. They wonder whether they will get used to these temperatures the next few weeks. If one compares the average loss of fluid in litres per day and the loss of salts in g salt/day of the last week for their visit to the first week, what is the most probable result?

- a. both fluid loss and salt loss will have decreased
- b. both fluid loss and salt loss will have increased
- c. fluid loss will have increased and salt loss will have decreased
- d. fluid loss will have decreased and salt loss will have increased.

These questions are, with some initial training, relatively easy to write, particularly because they come close to what clinicians do in actual clinical practice. The response format is a menu. The length of the menu does not need to be fixed, but is usually as long as there are meaningful alternatives.

Another MCQ type also proposed by the US National Board of Medical Examiners was Extended Matching Questions (EMQs). Originally this was introduced as a 'pattern recognition test' (Case & Swanson 1993, Case et al 1988). Students are presented with a series of brief case scenarios based on a single chief complaint (e.g. shortness of breath) and must select the most appropriate diagnosis or action from a menu of options. EMQs are relatively easy to construct.

MCQs of the kind described represent clinical reasoning formats in their simplest form. They are characterized by a professionally authentic stimulus format in combination with a closed response format. Reliability is similar to that of normal MCQs (Case et al 1994). Stimulus formats with richer (and longer) vignettes contain more 'measurement information' and contribute better to reliability than other vignettes. Longer menu response formats may appear to be better, but recent evidence suggests no advantage over simple 5-option MCQs (Swanson et al 2005). More complex response formats (e.g. using multiple best answers or allowing logical operators between different elements) and more complex scoring systems (like penalties and partial credit) are not recommended. Simple single best-answer formats and simple scoring systems are advised. In all, simple strategies seem to work best. An excellent manual for writing these MCQs is available (Case & Swanson 2002) and is freely available from the website of the US National Board of Medical Examiners (www.nbme.org).

OTHER CLINICAL REASONING FORMATS

On the basis of cognitive expertise theory, Charlin and his co-workers proposed the Script Concordance Test (SCT) (Charlin et al 2000). Most clinical problems are ill-defined, and experts do not collect exactly the same data and do not follow the same paths of thought. They also show substantial variation in performance on any particular real or simulated case. Their reasoning performance is based on illness scripts that have been shaped through individual training, experience and clinical exposure. Charlin et al challenged existing MCQ-based formats for their characteristic of applying well-known solutions to well-defined problems requiring a unique right solution. The SCT, in contrast, uses ill-defined problems and a method called aggregate scoring (Norman 1985) that takes expert variability into account. A clinical scenario is presented that provides a challenge to the candidate since not all data are provided for solution of the problem. A menu of options is presented from which the candidate may score the likelihood of each option in relation to the solution of the problem on a +2 to -2 scale. An example is:

A 25 year-old male patient is admitted to the emergency room after a fall from a motorcycle with a direct impact to the pubis. Vital signs are normal. The X-ray reveals a fracture of the pelvis with a disjunction of the pubic symphysis.

■ followed by a series of questions like:

<i>If you were thinking of</i>	<i>And then you find</i>	<i>This hypothesis becomes</i>
<i>Urethral rupture</i>	<i>Urethral bleeding</i>	<i>-2 -1 0 +1 +2</i>

The scoring reflects the variability experts demonstrate in the clinical reasoning process. Credits on each item are derived from the answers given by a reference panel. The credit for each answer is the number of reference panel members that have provided that answer, divided by the modal value for the item. For example, if on a particular item six panel members (out of 10) have chosen response +1, this choice receives 1 point (6/6), and if three experts have chosen response +2, this choice receives 0.5 (3/6). The total score for the test is the sum of credits obtained on all items.

Numerous studies of the validity of the SCT have been conducted (Charlin & van der Vleuten 2004). Reliability is quite good, showing that a value of 0.80 is reached with approximately 1 hour of testing using about 80 items.

OTHER CLINICAL REASONING ASSESSMENT METHODS

In the recent literature other methods have also been proposed. However, they either have had, as yet, less impact on the assessment field or are supported by only limited research into their measurement properties.

An instrument that has some resemblance to the SCT is a test called the Clinical Reasoning Problem (CRP) (Groves et al 2002). The CRP is intended specifically to assess the *process* of clinical reasoning, not so much the outcome. The stimulus format consists of a clinical scenario including a presentation, history and physical examination. Subjects are asked to nominate the two diagnoses they consider most likely, to list the features that they regard as important in formulating their diagnosis, to indicate whether these features are positively or

negatively predictive, and to give a weighting of each. There is not necessarily a single correct answer. Scoring is again done by using information from an expert panel. Reliability of the CRP seems comparable to that of MCQs and moderate correlations are found with criterion variables.

Finally, the Clinical Reasoning Exercise has been designed to assess students' knowledge of the basic mechanisms of disease (Neville et al 1996). The stimulus format presents short clinical presentations, with history and examination data as a stimulus format and a one-paragraph write-in answer as the response format. Approximately 15 cases are required for an acceptable level of reliability (0.78), and consistency of scores across multiple tests is excellent (0.84) (Wood et al 2000). Moderate correlations have been found with a knowledge test.

IMPLICATIONS AND ADVICE FOR THE TEACHER

As has become evident from this review, our success in developing valid measures of clinical reasoning for student assessment has been a sobering experience. Clearly *the* method for clinical reasoning assessment does not exist. It is clear that our intuitive notions of complex clinical simulations are not what we might have expected from them in the first place. Simpler simulation technologies, with capacity for much greater sampling, seem to do a better job. If this is the disheartening reality, what should we as educators do in day-to-day practice? Are there some guidelines that could be developed from the findings so far which would allow us to proceed with some forms of assessment of clinical reasoning, albeit with caution? Unfortunately there are no fixed answers to these questions. For instance, the answer may be quite different for tests which are to be used in undergraduate courses largely for formative purposes than for those used for major postgraduate certifying examinations where high levels of reliability are demanded.

There are several key points we wish to make. First, it is hard to imagine a credible assessment of clinical competence which does not attempt to evaluate clinical reasoning skills. An assessment using less-than-perfect instruments is preferable

to no assessment of this component at all. This is an issue of validity which must apply to the whole assessment procedure.

A second compelling argument against discarding our imperfect instruments is the very direct and powerful relationship between assessment and student learning. Academic success is largely defined by examination performance and academic success is what students are seeking. Thus, students will devote much of their energy to identifying and studying what they believe will be in their examination (van der Vleuten & Schuwirth 2005). This impact of examinations on student learning will often be greater than that of the training programme and is sometimes referred to as *consequential validity* (Messick 1995). Such effects must be seen as inevitable, if not desirable. The only answer is to ensure a good match, at least in students' minds, between the assessment procedures and the expected outcomes of the course. Failure to do so may have serious consequences. The bottom line is that a choice for a particular method may be motivated because of its (expected) education effect. For example, in a recent presentation on the assessment programme of a PBL school, the use of the somewhat older modified essay questions was maintained even at the cost of substantial resources because of the beneficial effect on the assessment of the learning of students (Prideaux 2006).

Finally, as this chapter makes clear, many ways to assess clinical reasoning are available and some are quite ingenious and creative. If no single measure is *the* measure, the choice is really yours. Which method appeals to you or your institution? How much effort do you wish to invest in writing simple or more complex stimulus formats? How many resources would you like to spend on the response format? What sort of reliability is required in your setting? What kind of impact do you strive for? What affinity or convention exists in your situation in relation to clinical reasoning assessment? Answers to these questions may vary considerably across different education contexts. A deliberate and motivated choice among the many possibilities that the literature now has to offer is on your agenda. The simpler your selected approach, the more you can rely on existing technologies and procedures and the less you will need to invest in unique solutions.

References

- Bransford J, Sherwood R, Vye N et al 1986 Teaching thinking and problem solving: research foundations. *American Psychologist* 41:1078–1089
- Brooks L R, Norman G R, Allen S W 1991 Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General* 120(3): 278–287
- Case S M, Swanson D B 1993 Extended-matching items: a practical alternative to free response questions. *Teaching and Learning in Medicine* 5:107–115
- Case S M, Swanson D B 2002 Constructing written test questions for the basic and clinical sciences. National Board of Medical Examiners, Philadelphia
- Case S M, Swanson D B, Stillman P S 1988 Evaluating diagnostic pattern recognition: the psychometric characteristics of a new item format. Paper presented at 27th Conference on Research in Medical Education, Washington DC
- Case S M, Swanson D B, Ripkey D R 1994 Comparison of items in five-option and extended-matching formats for assessment of diagnostic skills. *Academic Medicine* 69: S1–S3
- Charlin B, van der Vleuten C 2004 Standardized assessment of reasoning in contexts of uncertainty: the Script Concordance approach. *Evaluation and the Health Professions* 27:304–319
- Charlin B, Roy L, Brailovsky C et al 2000 The script concordance test: a tool to assess the reflective clinician. *Teaching and Learning in Medicine* 12:189–195
- De Bruin A B, Schmidt H G, Rikers R M 2005 The role of basic science knowledge and clinical knowledge in diagnostic reasoning: a structural equation modeling approach. *Academic Medicine* 80:765–773
- Elstein A S, Shulman L S, Sprafka S A 1978 *Medical problem solving: an analysis of clinical reasoning*. Harvard University Press, Cambridge, MA
- Eva K W 2005 What every teacher needs to know about clinical reasoning. *Medical Education* 39(1):98–106
- Farmer E A, Page G 2005 A practical guide to assessing clinical decision-making using the key features approach. *Medical Education* 39:1188–1194
- Feletti G, Engel C 1980 The modified essay question for testing problem-solving skills. *Medical Journal of Australia* 1:79–80
- Feltovich P J, Barrows H S 1984 Issues of generality in medical problem solving. In: Schmidt H G, De Volder M L (eds) *Tutorials in problem-based learning: a new direction in teaching the health professions*. Van Gorcum, Assen, p 128–142
- Fischer M R, Kopp V, Holzer M et al 2005 A modified electronic key feature examination for undergraduate medical students: validation threats and opportunities. *Medical Teacher* 27:450–455
- Groves M, Scott I, Alexander H 2002 Assessing clinical reasoning: a method to monitor its development in a PBL curriculum. *Medical Teacher* 24(5):507–515
- Hatala R, Norman G R 1999 Influence of a single example upon subsequent electrocardiogram interpretation. *Teaching and Learning in Medicine* 11:110–117
- Hatala R, Norman G R 2002 Adapting the key features examination for a clinical clerkship. *Medical Education* 36:160–165
- Mcguire C H, Babbott D 1967 Simulation technique in the measurement of problem-solving skills. *Journal of Educational Measurement* 4:1–10
- Messick S 1995 The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 23:13–23
- Neufeld V R, Norman G R (eds) 1985 *Assessing clinical competence*. Springer, New York
- Neufeld V R, Norman G R, Feightner J W et al 1981 Clinical problem-solving by medical students: a cross-sectional and longitudinal analysis. *Medical Education* 15(5): 315–322
- Neville A J, Cunnington J, Norman G 1996 Development of clinical reasoning exercises in a problem-based curriculum. *Academic Medicine* 71:S105–S107
- Norcini J J, Swanson D B, Grosso L J et al 1985 Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. *Medical Education* 19:238–247
- Norman G R 1985 Objective measurement of clinical performance. *Medical Education* 19:43–47
- Norman G 2005 Research in clinical reasoning: past history and current trends. *Medical Education* 39:418–427
- Norman G, Tugwell P, Feightner J et al 1985 Knowledge and clinical problem-solving. *Medical Education* 19:344–356
- Norman G R, Brooks L R, Allen S W 1989 Recall by experts and novices as a record of processing attention. *Journal of Experimental Psychology: Learning, Memory and Cognition* 5:1166–1174
- Norman G, Allery L, Berkson I et al 1992 Research in the psychology of clinical reasoning: implications for assessment. Cambridge Conference IV. Cambridge, Office of the Regius Professor, Cambridge University
- Norman G, Swanson D, Case S 1996 Conceptual and methodology issues in studies comparing assessment formats: issues in comparing item formats. *Teaching and Learning in Medicine* 8:208–216
- Norman G, Bordage G, Page G et al 2006 How specific is case specificity? *Medical Education* 40(7):618–623
- Page G, Bordage G 1995 The Medical Council of Canada's key features project: a more valid written examination of clinical decision-making skills. *Academic Medicine* 70:104–110
- Page G, Bordage G, Harasym P et al 1990 A new approach to assessing clinical problem-solving skills by written examination: conceptual basis and initial pilot test results. In: Bender W, Hiemstra R J, Scherpbier A et al (eds) *Teaching and assessing clinical competence*. Groningen, Boekwerk Publications, Groningen, The Netherlands, p 403–407

- Petrusa E R 2002 Clinical performance assessments. In: Norman G R, van der Vleuten C P M, Newble D I (eds) *International handbook for research in medical education*. Kluwer Academic Publisher, Dordrecht, p 673–709
- Prideaux D 2006 Constructed response items: MEQs & SAQs. IDEAL Train the Trainer Assessment Workshop. Muscat, Sultanate of Oman
- Regehr G, Norman G R 1996 Issues in cognitive psychology: implications for professional education. *Academic Medicine* 71(9):988–1000
- Schmidt H, Norman G, Boshuizen H 1990 A cognitive perspective on medical expertise: theory and implications. *Academic Medicine* 65:611–622
- Schuwirth L W T, van der Vleuten C P M, Donkers H H L M 1995 Computerized long-menu questions, an acceptable un-cue-version. In: Rothman A I, Cohen R (eds) *The sixth Ottawa Conference on Medical Education*. University of Toronto Bookstore Custom Publishing, Toronto, p 178–181
- Schuwirth L W T, van der Vleuten C P M, Donkers H H L M 1996a A closer look at cueing effects in multiple-choice questions. *Medical Education* 30:50–55
- Schuwirth L W T, van der Vleuten C P M, De Kock C A et al 1996b Computerized case-based testing: a modern method to assess clinical decision making. *Medical Teacher* 18:295–300
- Schuwirth L W T, Blackmore D E, Mom E et al 1999 How to write short cases for assessing problem-solving skills. *Medical Teacher* 21:144–150
- Schuwirth L W, Verheggen M M, van der Vleuten C P et al 2001 Do short cases elicit different thinking processes than factual knowledge questions do? *Medical Education* 35:348–356
- Skakun E N, Maguire T O, Cook D A 1994 Strategy choices in multiple-choice items. *Academic Medicine* 69: S7–S9
- 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* 12:220–246
- Swanson D B, Holtzman K A, Albee K et al 2005 Psychometric characteristics and response times for content-parallel extended-matching and one-best-answer items in relation to number of options. *Academic Medicine* 81(10 Suppl):S52–S55
- Turnbull J, Van Barneveld C 2002 Assessment of clinical performance: in-training evaluation. In: Norman G R, van der Vleuten C P M, Newble D I (eds) *International handbook of research in medical education*. Dordrecht, Kluwer Academic Publishers, p 793–810
- van der Vleuten C P M, Dolmans D H J M, Scherpbier A J J A 2000 The need for evidence in education. *Medical Teacher* 22:246–250
- van der Vleuten C P M, Schuwirth L W T 2005 Assessment of professional competence: from methods to programmes. *Medical Education* 39:309–317
- Wood T, Cunnington J, Norman G 2000 Assessing the measurement properties of a clinical reasoning exercise. *Teaching and Learning in Medicine* 12:196–200