Computerized case-based testing: a modern method to assess clinical decision making

L. W. T. SCHUWIRTH[1], C. P. M. VAN DER VLEUTEN[1], C. A. DE KOCK[2], A. G. W. PEPERKAMP[1] & H. H. L. M. DONKERS[3]

[1]*Department of Educational Research and* [2]*Department of General Practice,* [3]*Department of Medical Informatics, University of Limburg, Maastricht, The Netherlands*

SUMMARY *This article describes an assessment system that has been developed to focus on application of knowledge. Its two major features are case-based testing and the use of multimedia and computer tools. The case material for this testing type is based on the key-features concept, reporting the most relevant characteristics of a case and asking a limited number of questions, each aimed at essential decisions. These cases are produced in daily practice using real patients. Subsequently an extensive review process is used to check for flaws in description, phrasing or answer keys. Cases are stored in an item-bank out of which an automated random stratified selection can be drawn, according to a pre-specified blueprint. Cases are then presented to the students by way of a specially developed interface using multimedia when indicated. The article further describes advantages, disadvantages and experiences.*

Introduction

Assessment of clinical reasoning has attracted major attention as one of the components of measuring clinical competence (Van der Vleuten & Newble, 1995). In the 1960s, partly resulting from concern about the knowledge orientation of most examinations and the widespread use of multiple-choice questions (MCQs), instruments were proposed to measure clinical reasoning skills (Rimoldi, 1961; McGuire & Babbott, 1967, Helfer & Slater, 1971). These measures typically confront students with patient information, and require the examinee to collect further data and to take diagnostic and management decisions. The quality of the decisions and the pathway through the clinical problem were used to judge the clinical reasoning skills of the examinee. The most prominent exponent was the Patient Management Problem (PMP; McGuire & Babbott, 1967; McGuire & Solomon, 1976) which became popular in many medical schools and certifying agencies. Some types of PMPs used complex branching pathways, rather than merely allowing a fixed or linear route through the problem. Ingenious technical devices (e.g. latent imaging techniques to reveal invisible ink) were used to withhold data until selected by the examinee. Later, alternative approaches such as sequential management problems (Berner *et al.*, 1974); modified essay questions and computer simulations (Williams *et al.*, 1984; Norcini *et al.*, 1986) were proposed. The common denominator in all these instruments is the use of a patient simulation presenting an examinee with a realistic clinical problem. The examinee is required to work through the entire simulation in order to solve the problem.

Research data and practical experience revealed a number of problems with this type of assessment (Swanson *et al.*, 1987). First, the scoring was a concern. Apart from managing the complexity of some of the scoring systems, it appeared to be quite difficult to reach consensus on appropriate pathways, correctness of diagnoses, management decisions and weighting schemes for rewarding proficiency, thoroughness and overall competence. Second, it appeared that a score on one simulation was not predictive—did not correlate—with a score on another simulation (Norman *et al.*, 1985). Therefore, a problem of reliability occurred. This low correlation implies that if an examinee were submitted to a random set of different but parallel problems, a different score would be obtained. Consequently, many problems would be required to achieve a reproducible or reliable score, yielding an impractically long test. This was called the content-specificity problem of clinical reasoning. Finally, some findings cast serious doubts on the validity of these methods. The fact that in some studies medical students out-performed more experienced doctors (Goran *et al.*, 1973; Marshall, 1977; Newble *et al.*, 1982) seriously questions the construct validity. Furthermore, scores on clinical reasoning measures, if reliably measured, were found to correlate very highly with regular multiple-choice questions (Norcini *et al.*, 1983). This raises the question of whether these measures really tested a different trait than simple factual knowledge.

In recent years, evidence from cognitive psychological

Correspondence: L. W. T. Schuwirth, Department of Educational Research, University of Maastricht, PO Box 616, 6200 MD, Maastricht, The Netherlands. Email: l.schuwirth@educ2.unimaos.nl

0142-159X/96/040294-06

research on the development of medical expertise has generated new insights into the nature of clinical reasoning (Schmidt *et al.*, 1990). The ability to reason is no longer considered a generic trait that generalizes across situations, problems or time, with growth characterized as a monotonous process. Rather, expertise develops as a transition from a conceptually rich and rational knowledge base (acquired from educational experiences) to a non-analytical ability to recognize and handle situations efficiently (acquired from extensive clinical experience). The expert uses a sophisticated form of pattern recognition characterized by speed and efficient use of information. These cognitive processes are highly dependent on the clinical content. Expertise is not a trait but a 'state', depending on the specific problem and personal experience.

These findings had a number of implications for assessing clinical reasoning (Bordage & Page, 1987; McGuire, 1987; Swanson, 1987; Norman *et al.*, 1992; Swanson & Case, 1993; Van der Vleuten *et al.*, 1995; Newble *et al.*, 1995):

- To overcome the problem of content-specificity more content can be sampled if the problem given to the examinee is focused on critical elements only. Instead of working through each patient case extensively, only the key aspects are focused on instead and other aspects are skipped, yielding more time to test more problems. This has been called the 'Cambridge-case' or 'key-feature' approach (Bordage & Page, 1987).
- What is being measured is more dependent on the task given to the examinee than on the format used; i.e. the content of the test is more important than the characteristics of the measurement method. In this respect, a distinction must be made between stimulus (the task posed) and response format (MCQ, essay, etc.). The validity of the stimulus will primarily determine the validity of the test; the response format is of secondary importance.
- The stimulus format should be 'conceptually and clinically rich'. The problems given to the examinees should approximate the real world as closely as possible. Clinical details and patient context information (i.e. the way the person looks or walks) provides important triggers for approaching clinical problems.

In this theoretical context, the medical school of the University of Limburg developed an assessment method using these implications as guidelines. Furthermore, the introduction of the method was motivated by the school's wishes to reward higher order cognitive skills more in the assessment system of the problem-based curriculum and the desire to introduce more standardization in assessment during the clinical rotations in the last 2 years of the (6 year) program (Van der Vleuten & Swanson, 1990). The method has been called Computerized Problem-based Testing (CPT). In essence, CPT represents nothing new nor anything spectacular, but it merely synthesizes a number of the previous developments in case-based testing. This article describes the method in more detail.

Computerized case-based testing

CPT is a testing procedure using a set of programs implemented on a network of personal computers. It consists of a part that allows the students to log on to the system and to be tested in a particular discipline (student interface) and a part that, in an interactive way, allows the teachers to convert written test material into computer tests (teacher interface). The system presents to the (authorized) student a test consisting of a prespecified number of patient cases based on a discipline-specific blueprint. This blueprint specifies the number of cases within relevant content areas. Since the cases are randomly selected within these content areas, each student is given a different sample of cases. The set of programs uses an MS Windows environment. The case text and graphical or pictorial information (photographs, X-rays, CAT scans) are revealed in separate windows. A mouse click on an 'OK button' opens a third window with the question. Cases may be accompanied by additional multimedia possibilities: pictures, moving images and sounds. After the question is answered, another question may follow or new case information may be added. When a case is completed, the student cannot return to a previous case, since in randomly generated tests one case may reveal the answer to another case. All items are scored immediately. Case and item scores are filed, expressed as a proportion of correct scores (in percentages). Immediately after completion of the test, the total score is presented in combination with feedback on group performance. Other feedback (i.e. correct answers, study suggestions), although technically possible has not yet been realized since the primary purpose of the present test is summative.

The stimulus format

Type of cases. Cases used in CPT do not represent a real-time or a branched simulation but are, instead, in the key-feature format (Bordage & Page, 1987). A short description of the patient is given and only relevant characteristics, signs and symptoms are reported. A limited number of questions are asked, focusing on essential decisions that have to be made in that particular case. This makes it possible to present many different case per test, thus covering the domain optimally.

All cases are based on real-life patients in real practice. This guarantees that the cases are as realistic as possible. A case should have an adequate description allowing the examinee to answer the question correctly. This means not only that the amount of information given should be optimal, but also that this information should be stated unambiguously. Other possible sources of information, such as graphics, sounds and even short films, should be used only on indication. More detailed case development guidelines used are summarized in Table 1. An example of a case with a few questions is presented in Table 2.

Production of cases. All participating (practising) physicians first received training in which the purpose of CPT was explained, case-writing techniques were trained and common pitfalls of case and item writing were discussed. Subsequently they started writing up patient cases they

Table 1. Some guidelines for making case-based test-material

- Use real-life patients as a base for your cases
- Do not make your cases too long or too short. Always check whether all the information needed to answer the question(s) is available; eliminate redundant information
- Make sure that what you write down is a correct representation of all your findings. For instance, do not leave out important negative findings (like '*no* rebound tenderness')
- Do not make cases easier than real life. Use the most appropriate tool to present findings (e.g. use audio to present a heart murmur, instead of a written description)
- Do not make cases more difficult than real life. For instance, take into account that in real life most corrections can be done even after some days (e.g. a forgotten lab test can still be ordered after several days), whereas in a test situation no correction is possible after completion
- Use extra resources (pictures, sounds, moving images, etc.) only when indicated. The use of these tools when they are not needed creates an unwanted distraction
- Show all your material to colleagues for criticism, or re-read your material after a 'latency' period
- Questions should be aimed at decisions that have to be taken in this case. It is a waste of the authors' and students' time and energy to use a case description for the testing of general knowledge
- An examinee should not be able to answer the question without having read the case. This helps to ensure that the question is really about decisions in the case. Furthermore, the questions should be such that students could be allowed to use textbooks
- The question must be phrased correctly:
  —do not use semi-quantitative terminology (often, sometimes, usually, etc.)
  —use short sentences
  —focus on one issue per question
  —avoid absolute terminology (never, always, etc.)
- The correct answer must be defensibly correct; a false answer must be defensibly incorrect
- Choose the question type resembling real life the most (e.g. a multiple-choice type of question for lab tests, an open-ended question for a simple diagnosis, etc.)

normally encounter during their clinical consultations. This case material is submitted to a two-step review procedure. In the first step, a small committee (two people) reviews the test material for content and formulation. Critical comments are sent back to the authors with suggestions for alterations. This cycle may be repeated several times before the material is considered satisfactory. In a second step, the material is presented to a larger group (six to eight people) for a further review of the content of cases, questions and answers. Only after having passed this second review successfully is the case stored in an item-bank.

The response format

Questions linked to the cases are directed towards decisions rather than to (factual) knowledge, and are phrased in such a way that no disagreement about their meaning may arise.

So far, 12 different types of questions have been implemented. This enables the author to choose a question type most closely resembling real practice and to use optimally the potentials of the computer. Question types can be divided into multiple-choice type of questions, open-ended type of questions and 'compound' questions (using unrolling cases). Four types of multiple-choice questions are used (normal single-best answer multiple choice, MCQ; multiple choice allowing more correct options to be chosen, MMCQ; and true–false with and without a 'don't know' option). All these types are scored in a usual way. In the MMCQ a point is given for each correct decision (either to choose the option or to leave it open). The total points obtained are then divided by the number of options.

A special type of multiple choice is the Multiple Probability Estimate (MPE, Van Rossum *et al.*, 1990), in which a number of options are presented to the examinee. For each of the options the examinee has to judge the probability on a seven-point scale. To score performance an expert panel's score is used as a gold standard. Student scores are calculated based on the distance of the examinees' answers from the mean expert rating. An example of a case with an MPE is given in Figure 1.

Three types of open-ended questions are used. Two types use a dialogue box in which the examinee has to enter text only. The efficiency of the algorithm for scoring the answer, though, is too limited to be used in high-stakes examinations. Most algorithms for understanding natural language suffer from the same problem (Schuwirth *et al.*, 1992; Sabak, 1993). Based on the work of Veloski *et al.* (1993), two computerized forms of long-menu questions have been created, one for single answers and one for multiple answers. Both use the same principle: students type in their (short) answers in a dialogue box, and the program searches for 'hits' in an extremely long list (over 2500 alternatives). While typing, the examinee sees whether the answer has been recognized by the program. More information on these questions, including some of the characteristics and some psychometric qualities of these question types, has been reported elsewhere (Schuwirth *et al.*, 1994).

Finally three compound types of question types have been implemented. They all have in common that information to the examinee is provided in steps, and the same

Table 2. An example of a case with two questions

Mrs Van Horn, a 25-year-old waitress in a well-known restaurant, consults you because of abdominal complaints. She has had a piercing and cramping pain in the upper abdominal region for about two days. It radiates to her back. She also has a serious headache. She does not feel ill, has not had a fever. She is afraid to eat. Although she worked until 12.15 the night before she still slept very badly.
She would like you to prescribe something for her stomach because she has to go to a wine-tasting course in two days.
Further history taking reveals that she has never had such complaints before, she has neither heartburn nor eructation. She is somewhat nauseated but has not vomited. Her complaints do not correlate with the type of food she takes. There is neither diarrhoea nor constipation.
On physical examination you find nothing except extreme tenderness on palpation in the upper abdomen below the xyphoid.

Question 1: What lab-tests would you order?

You ordered the liver function tests plus amylase. Amylase was increased.

Question 2: This patient must be referred to an internist. True/false

You decide to refer her to an internist.

Question 3. Indicate the most appropriate level of urgency with which this patient must be seen by an internist.
a today
b this week
c this month

question is asked repeatedly. In one of these types, for example, the first part of a case reports the initial complaints. A question then asks for the options that should be included in the differential diagnosis. Further segments of the case then reveal the history, physical examination and further investigations. After each part the same question is repeated, but the number of correct options diminishes. Scoring in these question types is based on the proportion of correct answers given divided by the total number of options that had to be considered during the case. Other scoring options are currently under investigation.

The computer programs

Apart from the interface for the students presenting the cases and questions and recording answers, which has been described above, the set of programs contains an interface for the teacher and an item-bank managed by a database management system (DBMS) for storing and retrieving test material and students' scores. The teacher interface allows the teacher, in an interactive way, to convert all written test material to files that can be used by the computer. This involves application of the specially developed syntax, linking graphics to cases, etc. The student and teacher interfaces have been implemented, using Borland Pascal and Borland C + + for Windows. The item-bank is managed by Sybase SQL Server 10.0.1.

Why the use of computers? The decision to use the computer in CPT was made on the basis of a number of perceived advantages. Many of the resources normally used in examinations may be saved by using a computerized system for administration and scoring of tests, particularly when used in combination with automated storing and retrieving of test material. Ideally, these resources will be redirected towards maintaining or improving the quality of test-material. The use of multimedia tools can help to improve the fidelity of the cases. In written tests only pictures can be used, the reproduction of which is quite expensive. The computer may not only use moving images and sounds but can also reproduce images repeatedly without generating further costs.

Furthermore some typical advantages of computers can be used. First, cases providing information in steps, revealing the correct answer to the previous question prior to asking the next, may make both questions independent. In all types of written examinations this is not possible. Using computers, this is simply realized by preventing students from going back to a previous screen. Second, outcome variables other than the usual proficiency scores can be assessed including response time or number of corrections. This allows exploration of other measures of problem solving as suggested by Norman (Norman, 1988). Third, by the random selection of tests every student receives a different test, and cases (and scores) remain within the item-bank. This diminishes security threats.

Finally, the computer enables sequential testing. This is a testing procedure in which a limited number of cases are administered to all examinees. For those examinees whose scores are within a chosen interval from the cut-off score, an additional set of cases is presented. The others are excused from further testing. The number of cases, i.e. the length of the test, varies according to the ability of the candidate. The procedure optimizes reliability and efficiency.

On the other hand, the computer illiteracy of some of the students, the initial costs for software and hardware, the difficulty (or even impossibility) of scoring (long-answer) open-ended questions and the fact that taking an examination from a computer screen is more fatiguing as opposed to written examinations can be considered to be the main disadvantages.

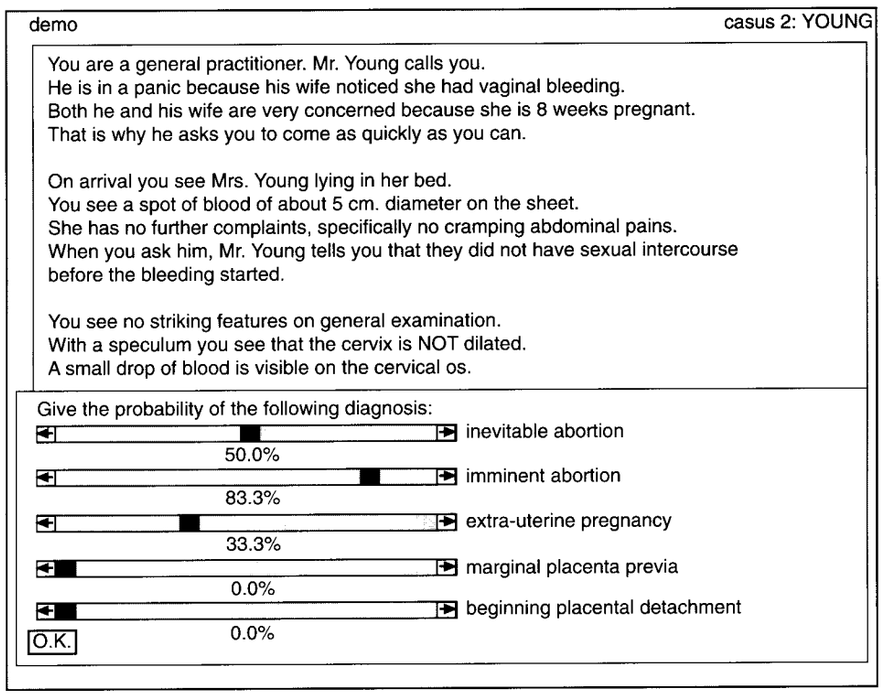

Figure 1. An example of a Multiple Probability Estimate question.
Note: The student uses the mouse to drag the button to the appropriate probability of each of the alternatives.

The experiences so far

Until now in only one discipline (general practice) have sufficient cases (about 500) been produced to enable the selection of a sample large enough to cover the domain adequately. In two other disciplines (gynaecology and internal medicine), the production of cases has started, but progresses only slowly. All authors indicate that the amount of time associated with production of high-quality cases is burdensome. Moreover, the number of corrections suggested by the review committee is substantial and requires considerable additional time. As experience of the case writers increases, however, the production time decreases to some extent. Production of good test material remains, nevertheless, a time-consuming activity. Despite this, case authors and other clinicians involved are very positive about CPT. They consider it a valuable additional tool to other, more knowledge-orientated, testing methods.

The limited psychometric information available indicates that scores from students at the end of a clinical rotation are well above chance level. The number of cases per hour of testing time is approximately 45. The reliability was modest (0.32 per hour of testing time), but may have been suppressed by the homogeneous nature of the group of students tested so far.

Student evaluations using an evaluation form with open-ended questions indicate that nearly all students like CPT. They perceive the content of CPT as highly relevant and adequately reflecting the practice in the rotation. Negative comments were related to the impossibility of correcting an answer on previous cases and the fact that working on computer screens is fatiguing.

So far, we are content with CPT. With further developments more research will be carried out. Not only should methods be found to improve reliability of the test scores (e.g. by using sequential testing methods) but also validity issues have to be addressed to assess whether using this method really has advantages over using more knowledge-based tests.

Notes on contributors

L. W. T. Schuwirth is an MD working in the area of test development and research for the medical faculty in Maastricht.

C. P. M van der Vleuten PhD is an educational psychologist, project leader of the project on assessment and evaluation of the medical faculty in Maastricht.

C. A. de Kock is a general practitioner participating in the project on development of the clerkship in general practice.

A. G. W. Peperkamp and H. H. L. M. Donkers are computer science engineers and programmers.

References

Berner, E.S., Hamilton, L.A. & Best, W.R. (1974) A new approach to evaluating problem solving in medical students, *Journal of Medical Education*, 49, pp. 666–672.

Bordage, G. & Page, G. (1987) An alternative approach to PMPs: the 'key features' concept, in: I.R. Hart & R.M. Harden (Eds) *Further Developments In Assessing Clinical Competence* (Montreal, Can-Heal).

Goran, M.J., Williamson, J.W. & Gonnella, J.S. (1973) The validity of patient management problems, *Journal of Medical Education*, 48, pp. 171–177.

Helfer, R.E. & Slater, C.H. (1971) Measuring the process of solving clinical diagnostic problems, *British Journal of Medical Education*, 5, pp. 48–52.

HODGKIN, K. & KNOX, J.D.E. (1975) *Problem-centered Learning* (London, Churchill Livingstone).

MARSHALL, J. (1977) Assessment of problem-solving ability, *Medical Education*, 11, pp. 329–334.

MCGUIRE, C. (1987) Written methods for assessing clinical competence, in: I.R. HART & R.M. HARDEN (Eds) *Further Developments in Assessing Clinical Competence*, pp. 46–58 (Montreal, Can-Heal).

MCGUIRE, C.H. & BABBOTT, D. (1967) Simulation technique in the measurement of problem solving skills, *Journal of Educational Measurement*, 4, pp. 1–10.

MCGUIRE, C.H. & SOLOMON, C.M. (1976) *Construction and Use of Written Simulations* (Chicago, Psychological Corporation).

NEWBLE, D.I., HOARE, J. & BAXTER, A. (1982) Patient management problems: issues of validity, *Medical Education*, 16, pp. 137–142.

NEWBLE, D.I., VAN DER VLEUTEN, C P.M. & NORMAN, G.R. (1995) Assessing clinical reasoning, in: J. HIGGS & M. JONES (Eds) *Clinical Reasoning Skills* (Oxford, Butterworth Heinemann).

NORCINI, J.J., MESKAUSKAS, J.A., LANGDON, L.O. & WEBSTER, G.D. (1986) An evaluation of a computer simulation in the assessment of physician competence, *Evaluation in the Health Professions*, 9, pp. 286–304.

NORCINI, J.J., SWANSON, D.B., WEBSTER, G.D. & GROSSO, L.J. (1983) A comparison of several methods of scoring patient management problems, in: *Proceedings of the 22nd Annual Conference on Research in Medical Education*, pp. 41–46 (Washington, DC, Association of American Medical Colleges).

NORMAN, G.R. (1988) Problem solving skills, solving problems and problem based learning, *Medical Education*, 22, pp. 279–286.

NORMAN, G., ALLERY, L., BERKSON, L., BORDAGE, G., COHEN, R., DAUPHINEE, D., DAVIS, W., FRIEDMAN, C., GRANT, J., LEAR, P., MORRIS, P. & VAN DER VLEUTEN, C. (1992) Research in the psychology of clinical reasoning: implications for assessment, in: *Proceedings of Cambridge Conference IV* (Cambridge, Office of the Regius Professor of Physics, Cambridge University School of Clinical Medicine).

NORMAN, G.R., TUGWELL, P., FEIGHTNER, J.W., MUZZIN, L.J. & JACOBY, L.L. (1985) Knowledge and clinical problem-solving, *Medical Education*, 19, pp. 344–356.

RIMOLDI, H.J.A. (1961) The test of diagnostic skills, *Journal of Medical Education*, 36, pp. 73–79.

SABAH, G. (1993) Knowledge representation and natural language understanding; *AICOM*, (3/4), pp. 155–186.

SCHMIDT, H.G., NORMAN, G.R. & BOSHUIZEN, H.P.A. (1990) A cognitive perspective on medical expertise: theory and implications, *Academic Medicine*, 65, pp. 611–621.

SCHUWIRTH, L.W.T., VAN DER VLEUTEN, C.P.M. & DONKERS, H.H.L.M. (1992) Het gebruik van open en gesloten vragen: Effecten van cueing en nauwkeurigheid van computergestuurde scoring [The use of open-ended and multiple-choice questions: effects of cueing and accuracy of computerized scoring], in: C.P.M. VAN DER VLEUTEN, A.J.J.A. SCHERPBIER & M.C. POLLEMANS (Eds) *Proceedings van het eerste Gezond Onderwijs Congres* (Houten/Zaventem, The Netherlands, Bohn Stafleu Van Lochum).

SCHUWIRTH, L.W.T., VAN DER VLEUTEN, C.P.M. & STOFFERS, J. (1994) Computerized long-menu questions, an acceptable un-cue version?, paper presented at the sixth Ottawa conference, Toronto, Canada.

SWANSON, D.B. (1987) A measurement framework for performance-based tests, in: I.R. HART & R.M. HARDEN (Eds) *Further Developments in Assessing Clinical Competence* (Montreal, Can-Heal).

SWANSON, D.B. & CASE, S.M. (1993) Extended-matching items: a practical alternative to free-response questions, *Teaching and Learning in Medicine*, (2), pp. 107–115.

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.

VAN DER VLEUTEN, C.P.M. & NEWBLE, D.I. (1995) How can clinical reasoning be tested?, *Lancet*, 345(8956), pp. 1032–1034.

VAN DER VLEUTEN, C.P.M. & SWANSON, D.B. (1990) Assessment of clinical skills with standardized patients: state of the art, *Teaching & Learning in Medicine*, 2, pp. 58–76.

VAN ROSSUM, H.J.M., BRIET, E., BENDER, W. & MEINDERS, A.E. (1990) The transfer effect of one single patient demonstration on diagnostic judgment of medical students: both better and worse, in: W. BENDER, R.J. HIEMSTRA, A.J.J.A. SCHERPBIER & R.P. ZWIERSTRA (Eds) *Teaching and Assessing Clinical Competence* (Groningen, The Netherlands, Boekwerk).

VELOSKI, J.J., RABINOWITZ, H.K. & ROBESON, M.R. (1993) A solution to the cueing effects of multiple choice questions: the Un-Q format, *Medical Education*, 27, pp. 371–375.

WILLIAMS, R.G., VU, N.V., BARROWS, H.G. & VERHULST, S. (1984) Profile of the Clinical Reasoning Test (CRT): an objective measure of problem solving skills and proficiency in using medical knowledge, in: H.G. SCHMIDT & M.L. DE VOLDER (Eds) *Tutorials in Problem-Based Learning (Assen, Van Gorcum).*