2009; 31: 759–763 MEDICAL TEACHER

Improving assessment with virtual patients

JONATHAN ROUND, EMILY CONRADI & TERRY POULTON
St. George's University of London, UK

Abstract

Assessments should accurately predict future performance in a wide variety of settings yet be feasible to conduct. In medical education a robust and comprehensive system of assessment is essential to protect the public from inadequate professionals. The parameters for devising such an assessment are well-defined, and good practice for writing examinations well-established. However even excellent written assessments are limited in their predictive validity, and limited in sampling, face and construct validity. The increasing availability and power of computing has led to growing interest in computer simulations for use in examinations, creating assessment virtual patients (AVPs). They can potentially test knowledge and data interpretation, incorporate images, sound or video and test decision making. Such AVPs could represent the most comprehensive, integrated assessment possible that is both objective and feasible. This article focuses on AVP design, distinguishing between linear and branched models, choice and consequence driven designs. It reviews the use of AVPs in the context of assessment theory. It presents different AVP designs discussing their benefits and problems. AVPs can become valuable components in high stakes medical exams, particularly in later years of courses. However this requires application of established assessment principles to AVP design.

Do we need assessment virtual patients?

Much development in medical education is driven by enthusiasts testing new tools. Computer-based assessment (CBA) has also been driven more by technological advancement than defined educational needs. Rightly Lambert Schuwirth asks "to which problem is [CBA] the solution?" (Schuwirth 2008). Assessment virtual patients (AVPs) are perhaps the most complex form of CBA, so should be subjected to careful scrutiny in their use, benefits and detractions.

Summative assessment in healthcare should determine if the candidate is ready to proceed in the course, or to start working autonomously. Society rightly expects educators to develop tools that identify those inadequate in knowledge, interpretation, application, skills, attitudes, communication and decision making. Most importantly assessment should predict future performance.

The tools available to the examiner make this difficult to achieve. The assessment must be objective and offer the same experience to all candidates, yet examine hard to measure skills, such as communication and physical examination, and do so in a relevant context, whilst not exposing patients to danger. It must test the whole curriculum, yet be feasible to deliver in terms of staff and cost. Ultimately, medical examination must attempt to determine if a candidate will perform adequately as a doctor, collect information, integrate and reference it against previous acquired knowledge and decide on a management plan. The characteristics of an ideal assessment are listed in Box 1.

Practice points

- Assessment Virtual Patients (AVPs) are computer-based simulations of patient management, incorporating narrative and media, designed to examine candidate skills in patient management.
- AVPs allow candidates to choose between different parts of history, examination, investigations or treatments. The clinical condition of the patient can depend on choices made.
- Scoring of AVPs can be based on choices made or the final condition of the 'patient'.
- AVPs offer many advantages over current assessment tools, offering integrated testing of knowledge, data interpretation and management skills set in a more realistic context than paper-based assessments.
- Clear understanding of what AVPs can test awaits carefully controlled experiments, but evaluations performed on AVPs in use have shown that they augment the existing range of assessment devices.

To answer Schuwirth, well-designed CBA offers possibilities for objective, integrated, comprehensive and context-appropriate testing of patient assessment and management. No other tool can do this, so CBA may be able to improve the quality of healthcare examination. Although still in its infancy, this area deserves proper development and evaluation.

Correspondence: Dr Jonathan Round, E-learning Unit, Centre for Medical and Healthcare Education, St. George's University of London, Blackshore Road, Tooting, London SW17 0RE, UK. Tel: +0208 725 2203; fax: 0208 725 0089; email: jround@sgul.ac.uk

ISSN 0142–159X print/ISSN 1466–187X online/09/080759–5 © 2009 Informa Healthcare Ltd.
DOI: 10.1080/01421590903134152

Box 1. Domains in an ideal assessment.

Content validity	Is it testing the intended subject area?
Concurrent validity	How does it compare with an established test?
Predictive validity	Will it predict future performance?
Construct validity	Is it able to test an abstract construct – decision making, empathy?
Face validity	Does it look (to the candidate) as if it is testing what it is supposed to?
Reliability	Will it consistently give the same result?
Feasibility	Can it be delivered technologically, logistically, and without excessive cost or staffing?

Computer-based assessment and assessment virtual patients

Since the 1960s computers have been used in medical examinations (Cantillon et al. 2004). At first this was essentially multiple choice questions (MCQs) or extended matching items delivered by computer. Initial concerns were that unfamiliarity with computers would affect performance although several studies, comparing modes of delivering a test, showed this did not occur (Lee & Weerakoon 2001). CBA offered advantages in cost, time and the potential creation of a large question bank that could even deliver a different test to each candidate, reducing security issues (Cantillon et al. 2004). Improved computing power has allowed images, sound and video to be incorporated, widening what can be tested, and the internet can allow distant participation. CBA has lagged behind the use of IT in teaching (Hols-Elders et al. 2008), perhaps because of the need for careful evaluation before use.

AVPs are one of the next developments in CBA, but represent a step change in what can be tested. They are computer based simulations of patient management designed to predict performance in clinical settings. They can also incorporate use of resources, including time or money. AVPs can be used synchronously or asynchronously, on or off line and can either produce marks visible or invisible to the candidate. They can be used in formative or summative assessments.

AVPs allow examiners to test knowledge, interpretation and decision making, all in the context of a patient scenario. Sound, images and video can be incorporated to test a candidate's skill in integrating multiple pieces of information into a presentation. Being multi-step, the patient can improve, deteriorate or develop new features. Because each candidate has the same 'patient', an AVP should be reliable. AVPs have excellent face validity and also would be expected to have better predictive validity than other formats, as no other test better integrates different patient management skills. AVP designs can incorporate consequences, based on choices made by the candidate. The exam can mirror real life, where a patient can be treated successfully in several different ways, and mistakes are often correctable.

There is concern that AVPs, being subject specific, reduce sampling and content validity. Also, although AVPs appear to operate in a 'virtually real' context, testing management, they are perhaps mostly testing knowledge (Schuwirth & van der Vleuten 2003). Increasing the number of AVPs within an exam and careful blueprinting reduces these issues. Indeed, Clauser found computer-based case simulations became a valuable component of the US medical licensing examination (USMLE) with just 2 h of testing time (Clauser et al. 2002).

There are several examples of AVPs being used in summative examinations. In Italy, 'computer-based case simulations' were introduced in 1999 to make up half of the National Medical Licensing Examination (Guagnano et al. 2002). AVPs have been used as OSCE stations in Sweden (Courteille et al. 2008) and Stanford University (Brutlag et al. 2006; http://websp.lime.ki.se/caseex), using adaptions of Web-SP. The USMLE Step 3 examination has used AVPs since 1999, and this is the most widely studied instance of AVP use (Dillon et al. 2004; http://www.usmle.org/). Step 3 tests complete patient management, so the candidate can request a large number of clinical and laboratory tests and order treatments in an uncued format. Responses are judged against the performance of experts in the field, and over-investigation or over-examination is adversely marked.

Structural design within assessment virtual patients

'Level 1' AVPs are essentially a series of MCQs based around a patient's presentation, assessment and management. Figure 1 shows a patient map, with a playable example at http://labyrinth.sgul.ac.uk/openlabyrinth/mnode.asp?id=qf4jesnqdknam1rx7jzarsx9qarsx9q. Correct responses send the user up the figure and incorrect ones down. As the patient 'progresses', the user moves from left to right. This AVP is scored by the final end point reached, although there may be several different ways to arrive there.

This type of AVP is extremely easy to build and use – typically each takes 1–2 h to create. It allows incorporation of sound, images and video. All candidates will get exactly the same experience, and it will have excellent face validity – it patently is testing the knowledge, understanding and data interpretation required to manage a patient. Despite its appearance in the figure, it is functionally linear, as choices made do not influence the condition of the patient or even the questions asked.

Similar AVP designs are used in the Swedish, Stanford and Italian AVPs, although with a much larger number of options and multiple combinations are selectable at each stage.

Assessments should separate candidates of different abilities, and Level 2 AVPs can be designed to do this better using adaptive testing (Cantillon et al. 2004). Although adaptive testing ranks candidates using fewer questions and less testing time (Kreiter et al. 1999), it has yet to be incorporated into high stakes healthcare examinations. However, it is relatively easy to design AVPs with adaptive testing. By adjusting the difficulty of the questions up and down the patient map (Figure 2), those who have done well in initial questions are asked harder ones at the end of the AVP and vice versa. This will help distinguish those who are poor from the mediocre and the excellent from the good.

None of these approaches really mimic real patient management, as choices made do not influence the 'patient'. Level 3 AVPs, incorporating choices and consequences allow

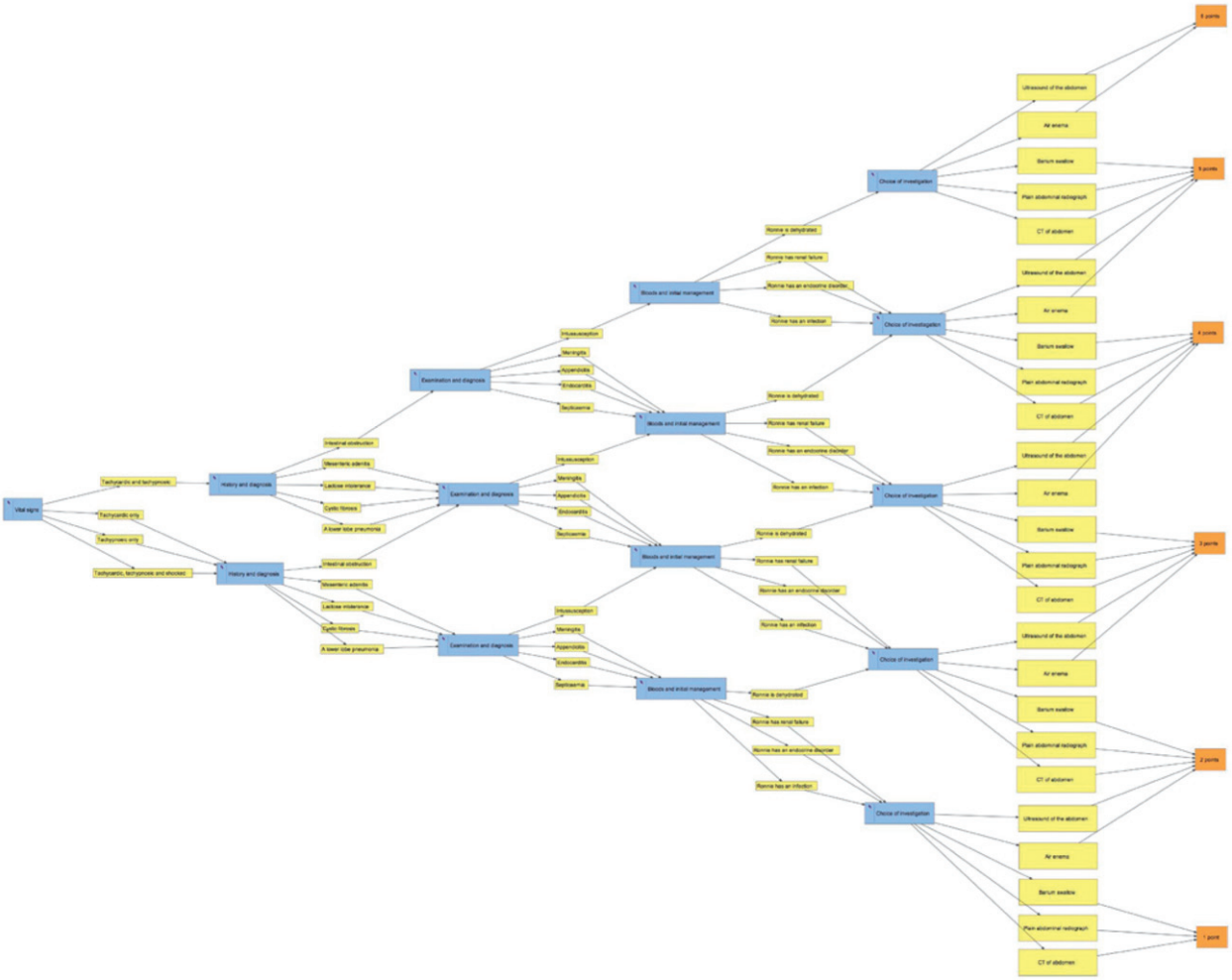

Figure 1. Level 1 AVP design.

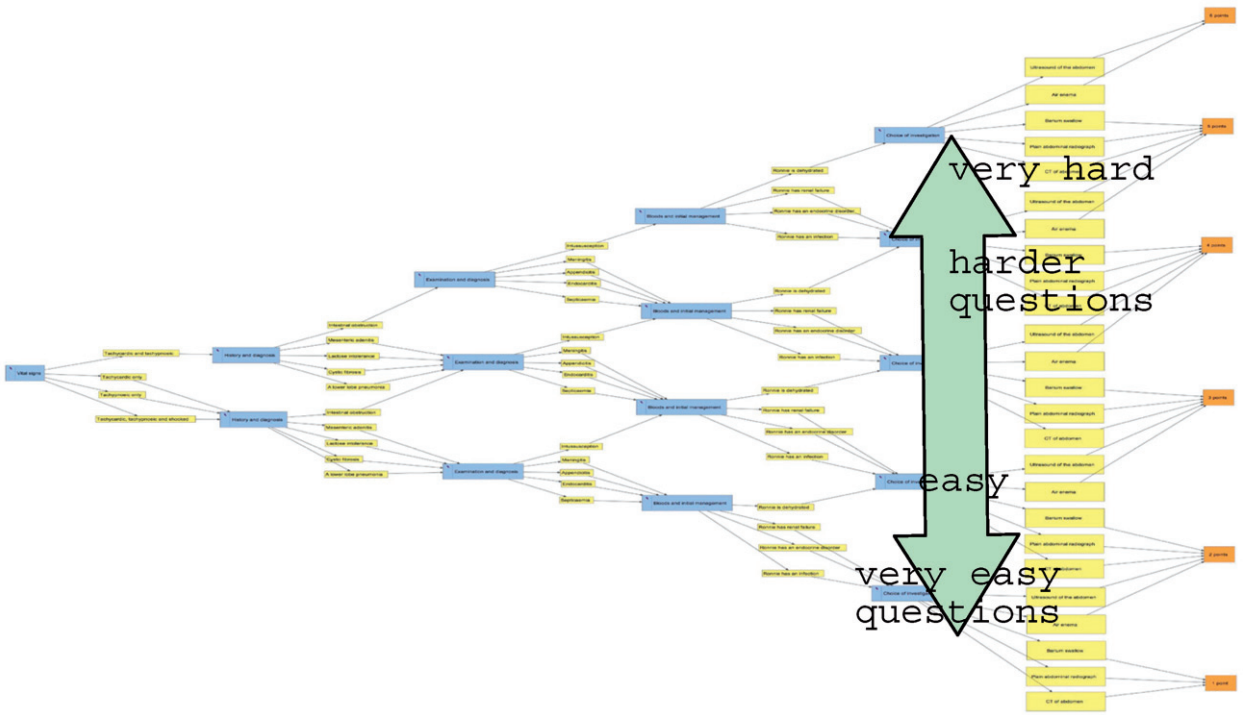

Figure 2. Level 2 AVP with adaptive testing.

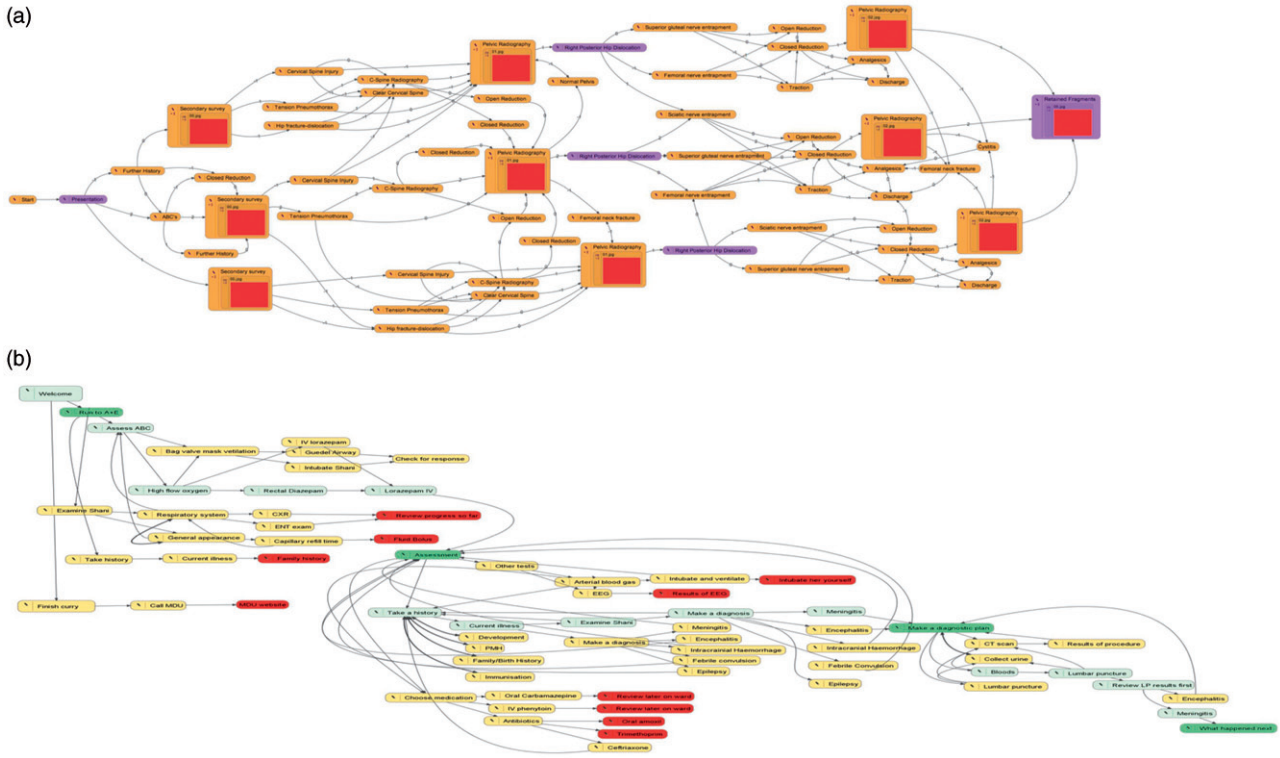

Figure 3. Level 3 AVPs. (a) Score dependent on routes taken and (b) score dependent on final point reached.

the candidate to correctly, or incorrectly, manage the AVP, with consequent change in the 'patient's' condition. Candidates can even identify and rectify poor decisions. Scoring can either relate to the decisions made, as shown in the patient map in Figure 3a or the final destination (Figure 3b), or both.

Level 3 AVPs (maps in Figure 3a and b) require typically 6–20 h per VP to create. Making those that address specific areas in curricula at an appropriate level is complex. Developing a bank of branched AVPs is taxing, but unless there were many AVPs in a single assessment, sampling would be poor.

Although closely reflecting the work of health care professionals, candidates will take different paths through the AVP, so the examination will not be the same for all. This could make assessment less reliable.

The USMLE uses a Level 3 model in Step 3, which feels very realistic, and scoring is based on choices made. However the effort needed to develop a bank of AVPs and the time needed for each case limits the sampling. In the USMLE, candidates 'treat' 9 patients in 4 h (http://www.usmle.org/).

Evaluation of assessment virtual patient use

Many thousands of candidates have now sat for examinations incorporating AVPs, but there is relatively little published evaluation of this assessment tool. However, it is becoming clear they test something distinct from other tools.

Qualitative data from pilot projects and surveys have explored the candidate experience. Candidates quickly work out how to use the AVPs and complaints directed at the software are rare (Conradi & Round 2008). Most negative comments relate to the candidate being shepherded in a particular direction, not being able to collect all the information wanted (Courteille et al. 2008), or select a particular option (Conradi & Round 2008). Positive comments focused on the realism, media usage and that actions affected outcome. AVPs were perceived to be fair, more so than MCQs, although no preference was shown over OSCE stations.

AVP scores correlate well with scores for other tests of knowledge. USMLE Step 3 performance (AVPs and 'understanding' MCQs) correlated well with Step 1 and Step 2 performance ('knowledge') (Andriole et al. 2005), although the authors were not able to separate AVP and MCQ scores. Sawhill found that, after correcting for Step 1 and Step 2 scores, Step 3 score improved in proportion with the length of training in a generalist training (e.g. medicine), but not with more focused programmes (e.g pathology) (Sawhill et al. 2003). AVP scores were correlated with length of time developing general patient management skills.

Researchers have also asked what AVPs are actually testing (Schuwirth & van der Vleuten 2003). Although attempting to test something beyond knowledge, candidates must know the subject area in order to demonstrate higher-order skills (decision making, prioritization) so scores will be domain specific (Schuwirth & van der Vleuten 2003). However, analysis of scores in different sections of the USMLE Step 3 (Clauser et al. 2002) and the Italian licensing exam (Guagnano et al. 2002) demonstrate poor correlation of AVP scores with tests of knowledge. Without a carefully controlled experiment designed purely to separate knowledge,

understanding and patient management it will not be clear what AVPs actually test.

Summary

Currently used exam formats struggle to objectively and reliably test knowledge, data interpretation, integration and patient management in a context representative of the work of a doctor. AVPs are computer delivered interactive patient scenarios where candidates must decide on management based on clinical information acquired during the course of the scenario. Candidates are scored based on the choices made or the final outcome. Different levels of complexity in AVP design are recognized. Potentially they are able to offer a better test of clinical abilities than other exam formats.

AVps take longer to create than single best answer or other formats, and require simple software to run. Concerns that they are really testing knowledge can be addressed by using many AVPS to sample different areas of the curriculum.

Despite being a major component of two high stakes exams worldwide, little is known of how they perform in practice. Further evaluation is needed to learn more about this new examination tool.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

Notes on contributors

JONATHAN ROUND is a Consultant Paediatrician and Senior Lecturer who runs the Child Health component of the Graduate Entry Programme at SGUL. His area of specialization is is Paediatric Intensive Care. His researc area is intensive care in oncology patients. Jonathan works with the e-Learning unit in the development of virtual patients and their rational use.

EMILY CONRADI works as an e-Projects Manager at SGUL. Emily has been involved with the PREVIEW and eViP projects as well as leading the development of the virtual patient design and development course.

TERRY POULTON is the Head of the e-Learning Unit at the Department of Medical Education at SGUL and responsible for both the delivery of e-Learning and all R&D. Terry has been responsible for development of the scenario based 'learning week' for the MBBS medical curriculum and the case-based learning Paramedic Foundation degree at SGUL.

References

Andriole DA, Jeffe DB, Hageman HL, Whelan AJ. 2005. What predicts USMLE step 3 performance? Acad Med 80(Suppl 1):S21–24.

Brutlag P, Youngblood P, Ekorn E, Zary N, Fors U, Gesundheit N. 2006. Case-ex: Examining the applicability of web-based simulated patients for assessment in medical education. World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (ELEARN).

Cantillon P, Irish B, Sales D. 2004. Using computers for assessment in medicine. BMJ 329(7466):606–609.

Clauser BE, Margolis MJ, Swanson DB. 2002. An examination of the contribution of computer-based case simulations to the USMLE step 3 examination. Acad Med 77(10):S80–S82.

Conradi E, Round J. 2008 Virtual patients as a tool for assessment Medbiquitous conference 2008 abstracts. Available from: http://www.ctsnet.org/abstractsprogram/medbiq/1802?view=displaymedbiqabstract&abstractid=31814

Courteille O, Bergin R, Stockeld D, Ponzer S, Fors U. 2008. The use of a virtual patient case in an OSCE-based exam – A pilot study. Med Teac 30:e66–e76.

Dillon GF, Boulet JR, Hawkins RE, Swanson DB. 2004. Simulations in the United States Medical Licensing Examination (USMLE). Qual Saf Health Care 13(Suppl 1):i41–i45.

Guagnano MT, Merlitti D, Manigrasso MR, Pace-Palitti V, Sensi S. 2002. New medical licensing examination using computer-based case simulations and standardized patients. Acad Med 77(1):87–90.

Hols-Elders W, Bloemendaal P, Bos N, Quaak M, Sijstermans R, De Jong P. 2008. Twelve tips for computer-based assessment in medical education. Med Teach 30(7):673–678.

Kreiter CD, Ferguson K, Gruppen LD. 1999. Evaluating the usefulness of a computerized adaptive testing for medical in-course assessment. Acad Med 74:1125–1128.

Lee G, Weerakoon P. 2001. The role of computer-aided assessment in health professional education: A comparison of student performance in computer-based and paper-and-pen multiple-choice tests. Med Teach 23(2):152–157.

Sawhill AJ, Dillon GF, Ripkey DR. 2003. The impact of postgraduate training and timing on USMLE step 3 performance. Acad Med 78(Suppl 1):S10–S12.

Schuwirth L. 2008. The use of computer-based assessment. Med Teach 30:651.

Schuwirth LW, van der Vleuten CP. 2003. The use of simulations in assessment. Med Educ 37(Suppl 1):65–71.

http://websp.lime.ki.se/caseex (Accessed 10 April 2009).

http://www.usmle.org/Examinations/step3/step3.html (Accessed 10 April 2009).