AN EVALUATION OF MEDICAL STUDENT ASSESSMENT USING VIRTUAL PATIENTS

Sophie Vaughan, Tristan Bate, Emily Conradi,
and Jonathan Round
Department of Medical Education
St George's, University of London
UK

Abstract

Virtual patients are interactive computer-based applications designed to mimic aspects of patient management. They can be used in assessment of students and professionals, but their use is neither widespread nor well researched. This project aimed to evaluate aspects of assessment virtual patient (AVP) design by creating a suite of AVPs and testing them with medical students. Analysis of performance was complemented with focus groups. Students completed the six cases taking 5–22 minutes for each one and scoring 75–100%. Feedback was very positive as students felt the cases tested their knowledge and management abilities well. A branched design was deemed more appropriate for final years and a linear approach for less experienced students. With these findings we are now developing cases for incorporation into summative final examinations.

Introduction

Assessment of doctors in training, in common with many other professions, is a closely monitored and tightly regulated process. The public expects that they will be seen by knowledgeable and technically competent doctors when they become ill. At the same time, the individual trainees expect that the assessment process is fair, predictable and impartial. The characteristics of an ideal assessment have been well described and are presented in Table 1.

Table 1: Characteristics of a 'high stakes' assessment (Dent & Harden 2005)

Content Validity	Are you testing what you think you are testing?
Predictive Validity	Does the test predict future performance?
Construct Validity	Can the test assess an abstract construct, such as empathy or decision making?
Face Validity	Does the test appear to the candidate to assess what it is meant to?
Reliability	Can the test produce similar marks/ranking each time?
Feasibility	Can the assessment be put on by the institution?
Safety	The test does not endanger participants

These conflicting demands have led to a plethora of different tools used in the assessment of the various traits essential in a doctor. A trainee would now expect to complete single best answers, extended matching items, objective structured clinical examinations, short answer questions, write case reports, undergo case based discussion, complete a professional development portfolio and many other forms of assessment. A prominent trend has been the attempt to separate the assessment of different attributes in an attempt to improve assessment of each one. Simultaneously attempts have been made to make uniform the examination experience of each student to improve reliability of the assessment. However these developments have divorced the assessment from clinical reality.

The Development of Computer-based Assessment

Over the last 40 years there has been a dramatic development of information technology transforming most aspects of the workplace and home. However these changes have not been matched in medical trainee assessment (Hols-Elders et al., 2008). Almost all examinations are paper or personnel based with a high demand for organisation, synchronisation and co-ordination. It is unclear why technology-based systems have not been developed more in medical exams, although there are concerns over security, system reliability and resources (Cantillon, Irish, & Sales, 2004).

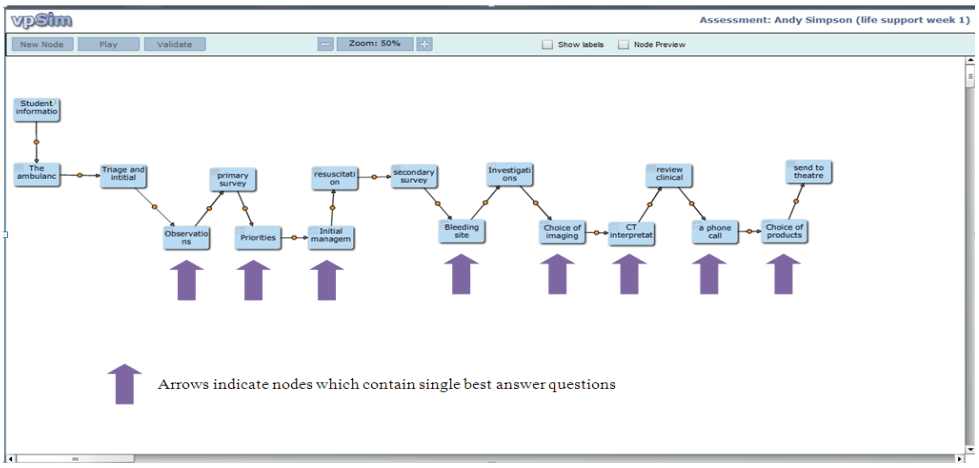
Despite this, simple computer based assessment is gradually becoming incorporated into medical schools. Initially this has been like-for-like replacement of paper based examinations. Some tools go beyond this, for instance allowing individualised feedback on student performance in formative assessments. Others allow examinations tailored in difficulty to students (Cantillon et al., 2004). The constraints of medical assessment are such that an ideal assessment is perhaps an impossible goal. However information technologies offer scope for repeatable patient based simulations that would be able, without inordinate cost, to examine the performance of trainees in a variety of clinical scenarios.

Types of Assessment Virtual Patients

Assessment virtual patients are interactive patient simulations designed for the purpose of distinguishing candidates of differing abilities. There are several different types in use for summative and formative assessment.

Level 1 AVPs are linear in design so have only one patient pathway (Figure 1.)

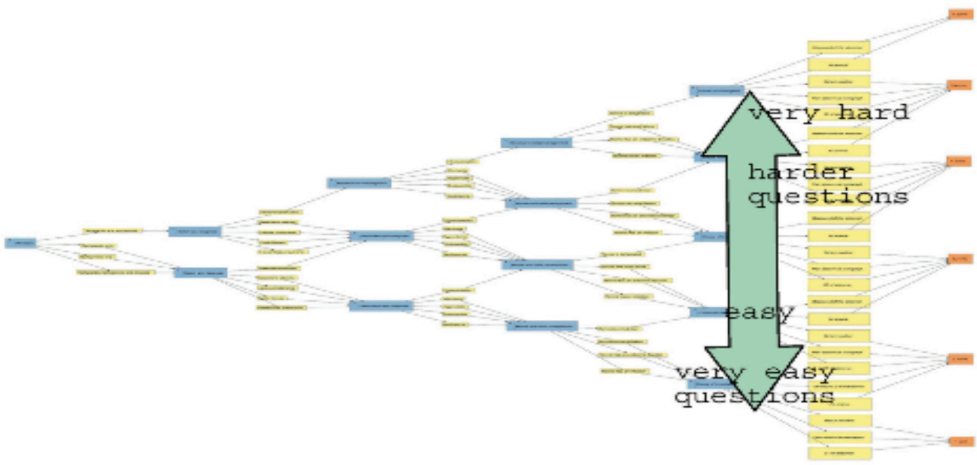
Figure 1: Screenshot of Level 1 AVP design used in testing



The patient's clinical condition is not determined by choices made. Single best answer questions (SBAs) along the pathway, relating to the patient presentation, physiology or management test the candidate's knowledge. A playable version can be found at: http://labyrinth.sgul.ac.uk/openlabyrinth/mnode.asp?id=qf4jesnqdknam1rx7j zarsx9qarsx9q.

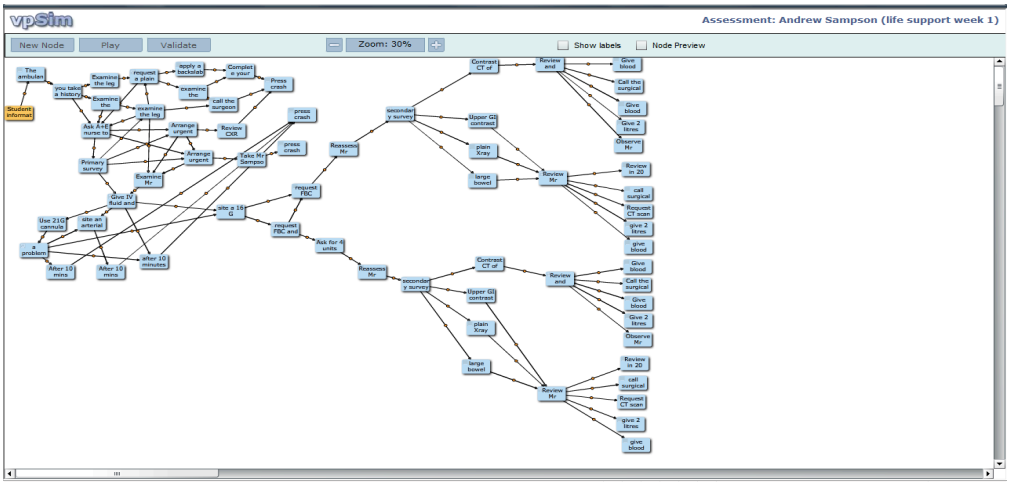
Level 2 AVPs are also linear in that, as for level 1 designs, the clinical condition of the patient is not affected by the choices of the examinee. However, the questions that the candidate has to answer differ depending on previous responses. Candidates answering initial questions correctly will get progressively harder questions to answer as shown in Figure 2. This process is termed adaptive testing, and it has been shown to discriminate between candidates right across the ability spectrum more quickly than non-adaptive tests (Kreiter, Ferguson, & Gruppen, 1999).

Figure 2: Level 2 AVP design



Level 3 AVPs are branched in design with multiple patient pathways as shown in Figure 3.

Figure 3: Screenshot of branched AVP



The clinical condition of the patient alters depending on the responses of the candidate to management choices. This attempts to recreate a life-like clinical scenario, wherein decisions made by the practitioner will affect the information gathered, tests performed, treatment given and outcome for the patient.

Scoring can be calculated from the eventual outcome or from the choices made. An example can be seen in Figure 3, and a playable version can be accessed at http://www.usmle.org/.

Assessment Virtual Patients

Assessment virtual patients, to the authors' knowledge, are only in use in four summative settings worldwide at the time of publication. In Italy, "computer-based case simulations" were introduced into the National Medical Licensing Examination in 1999 (Guagnano, Merlitti, Manigrasso, Pace-Palitti, & Sensi, 2002). AVPs are also used in OSCE stations in Sweden (Coureille, Bergin, Stockeld, Ponzer, & Fors, 2008) and at Stanford University (Brutlag et al., 2006). All of these instances share level 1 design features with an extended array of potentially correct options reflecting clinical choices.

The USMLE step 3 employs AVPs with a level 3 design approach, with intricate patient modelling to mimic patient management. Candidates must make their choices in uncued response boxes and scoring reflects not only the final condition of the patient, but time taken and financial cost.

Despite their important role in these settings, there is little published analysis of AVPs in terms of their usability, performance as an assessment tool or design features. What is known is that they are respected by students as easy to use with good face validity

(Conradi & Round 2008), and that scores correlate well with other tests of knowledge (Andriole, Jeffe, Hageman, & Whelan, 2005). More significantly, scores relate well to length of generalist training (Sawhill, Dillon, & Ripkey, 2003) suggesting that AVPs may be testing some aspect of clinical practise beyond knowledge.

Aims and Purpose of Evaluation

Virtual patients are expensive and time consuming to create, with estimates ranging from 8-30 hours even for a simple case (Huwendiek, Balasubramaniam, & Round, 2008). AVPs have additional problems in that they need to be of a high standard, free of glitches, secure and planned to test candidates at an appropriate level.

Before integrating AVPs into the assessment strategy of the medical school, we aimed to find out more about the performance of AVPs, and what design features were useful or problematic. By writing cases with contrasting design features and testing them in a formative assessment, we would be able to measure time taken, difficulty and paths through the cases to see how students used them. Using focus groups we would be able to examine the experience of students using them.

Methods

The cases were created using vpSim (vpsim.pitt.edu), an online virtual patient case-based authoring and playback system. VpSim provided the ability to create branched cases where decision points led to consequences affecting the patient outcome or single best answer questions within the patient pathway.

Participants

We tested the cases with two groups of medical students — pre-clinical and final year clinical students as shown in Table 2. The students were invited by email to attend a testing day during which they played each of the different AVP cases.

Table 2: Characteristics of participants from each focus group

Focus Group	Number of participants	Year of study
1	7	Final year
2	5	T year (2^{nd} to 3^{rd} years)

Assessment Virtual Patient Cases

The patient cases were a mix of emergency and clinic based cases. All were ‘first person players’. By example, candidates assumed the role of the junior doctor in the Emergency Department managing a young man following a road traffic accident and played as a GP managing a pregnant woman and subsequently her baby.

The AVPs were developed along two contrasting designs. One featured linear patient cases (e.g. Figure 1), where students were led through clinical cases while being tested with single best answer questions on their understanding of the clinical information and management. The other design was branched (e.g. Figure 3), with branching decision points such that their patient would improve, worsen or even die. The case features are summarised in Table 3.

Table 3: Description of the AVP suite

Case name	Andy Simpson	Joan - linear	Stephenson – linear	Andrew Sampson	Joan - branched	Stephenson - branched
Subject area	Trauma, shock	Cardiac, atrial flutter	Prenatal care, depression, growth	Trauma, shock	Cardiac, atrial flutter	Prenatal care, depression, growth
Case design	Linear	Linear	Linear	Branched	Branched	Branched
Pages	17	14	27	68	38	57
Narrative only	9	2	15	47	11	42
SBA questions	8	12	12	-	-	-
Branch points	-	-	-	21	27	15
Scoring	Visable postitive	Visable Negative	Visable Positive and negative	Hidden final score	Visable Negative	No scoring
Maximum possible score	40	100	120	10	100	-

Cases had different methods of scoring and feedback. These ranged from a mixture of positive and negative scoring which was either visible throughout the case, or hidden until a final management score was given at the end of the case. The feedback was immediate for questions and decisions during the linear cases, but was more subtle in the branched cases as the narrative and outcome would reflect the decision made.

Analysis

We collected data on the scores achieved, number of clicks and the time taken to play each case. After playing all the cases, the two groups of students participated in focus groups facilitated by a member of the medical education department. These evaluated and compared the linear and branched designs as well, different methods of scoring, feedback and usability. Tape recordings of both groups were transcribed and individual remarks coded by one author (TB).

Results

Analysis of the performance of the students undertaking the suite of virtual patients is shown in Table 4.

Table 4: Performance characteristics for the suite of AVP cases

Case name	Andy Simpson	Joan - linear	Stephenson - linear	Andrew Sampson	Joan - branched	Stephenson - branched
Case design	Linear	Linear	Linear	Branched	Branched	Branched
Max. poss. score	40	100	120	10	100	-
Score range	32-40	91-100	76-118	6-10	97-100	-
Mean score (% of max)	35 (88%)	97 (97%)	88 (73%)	8 (80%)	99 (99%)	-
Time taken mins (range)	9-22	9-15	12-19	5-12	6-10	9-13
Time taken mins (mean)	13	13	16	7	7	11

Focus Groups

The coding procedure applied to the focus group transcripts revealed six separate themes: value; skills/knowledge to be assessed; learning background; position in course; format; feedback. Analysis of comments is presented under these subheadings.

Value. Participants were overwhelmingly positive about the concept of using virtual patients as learning tools, stating that they were "really useful learning tools actually, even in this assessment format; really good for formative assessment" and they "would use [AVPs] over a book." Students imagined they might supplement their learning styles: "if you're on a GP rotation you've got a three hour lunch break. . .it would be really good to [use virtual patients]" and some "would [use AVPs for] revision and maybe self directed learning."

Most found the use of the dynamic stories in AVPs attractive: "I really enjoyed reading all the stories actually" but one student found narrative could be excessive: "the Stevenson branched one had an awful lot of pages that you're just clicking through and reading, and if I can't click back I find it too much to read."

Realism was also seen as a positive feature: "branched [was my preference] because it's more realistic." and the ability to rectify a non fatal error was seen as important: "with the branch ones you don't get a chance to correct yourself and that's quite important to me."

Skills/knowledge to be assessed. Several participants felt strongly that linear AVPs were more suited to testing basic sciences, "I think linear's better for basic knowledge, and branching for clinical decisions" and that "the multiple choice questions are particularly good for testing your knowledge of physiology or. . .if you understand how the mechanism works." Conversely students thought a branching design better for assessing decision making: "the branching format is good, in the trauma case that seemed really appropriate because it's about management decisions", although some questioned the restrictions imposed by the format: "there is room for perhaps doing things in a slightly different order. So that seemed particularly good for [testing management decisions]."

As well as testing decision making, one student felt that branched AVPs tested understanding on a deeper level: "I found the linear ones were about revision, so you could still do them just to make sure you remembered it. The branched ones really made me think, made me slow down and stop and think about stuff", and reflected the complexities of real life: ". . .if there's a less definitive answer, then [branched AVPs are] much more appropriate." The possibility of having a hybrid format, incorporating elements of both linear and branched styles, was also raised.

Learning background. Some students commented that their preferred type of AVP was influenced by the teaching methods and curriculum on their particular course. Those that had a background in basic sciences or more traditional teaching methods seemed to prefer the clinical realism of branched AVPs:

> I like the decision making. . .moving on to management now it works quite well. . .it's really good to get the branch cases and have the options of doing it more on a clinical level as opposed to just basic sciences. (3rd year student who studied basic medical sciences at Cambridge)

On the other hand, one Graduate Entry Programme (GEP) student who was more familiar with case based learning indicated a preference for a linear structure:

> From the start we've been taught [on the GEP course], from our very first case, from a clinical point of view and I just feel that my worries are about not having enough basic science. (2nd year GEP student, indicating a preference for linear AVPs)

Position in course. The relevance of the position of students in their course emerged as a strong theme from the focus groups. Senior students valued the opportunity to use AVPs, but also felt they would be applicable to junior students, with questions tailored to their stage of training:

> there is a big potential [for AVP use] for the first and second years... rather than saying which drug would you give [the patient], the doctor has decided to give them Dopamine, what kind of drug is this?. . .they want to get into the clinical part of [medicine].

Many participants indicated that they felt linear AVPs were suited to junior students, with branched AVPs appropriate for senior students with more developed clinical reasoning:

> I can see the linear one being useful for the preclinical students, when they're still learning their basic sciences. Now we're moving into clinical sciences, clinical management I think it should be more about application of knowledge.

A final year student stated that he "would prefer branched AVPs so long as you get the same level of feedback about why you're right or why you're wrong."

Format. Some of the cases taken by the students incorporated photographs or scanned clinical resources such as growth charts or ECGs. Students generally found them helpful saying that they "help you realise that you've switched characters. You're in hospital, and now you're back at the GP surgery and now you're an SHO." Others found that "it made [the cases] interesting" or that "you can remember the patient."

Time management was seen as important, especially if AVPs are to be used as summative assessments. One wondered "if you could number the questions or tell people at the beginning how many questions there are. . .so that you don't spend too much on doing one section?" and another suggested "a percentage bar to show how far along you are."

Feedback. Receiving feedback was a key issue for students in both focus groups: "Because you just don't get much through med-school. . .you rarely get told you're doing right." Most instinctively preferred the instantaneous feedback "The linear system is really nice because you know whether you've got it right or wrong and you can put it out of your mind." and found the lack of this a disadvantage with branched AVPs: "[with branched AVPs] you're feeling your way in the dark, you don't get that reassurance after each decision that you make".

Different mechanisms of feedback for branched AVPs were suggested: "you should show people the maps after they've gone through the VPs, [and] what direction it was supposed to go", or "your path. . .highlighted". Others suggested "a little score up in the corner. . .to know where you went wrong as you're going."

Students generally disliked negative marking, where it was only possible to lose marks rather than gain them, although one student found this method helpful. One student felt that allocating variable marks was a good scoring method for branched AVPs.

Conclusions

The primary purpose of this project was to understand how AVPs could best be developed as a large testing format for health care professionals. We had sought some direction as to length of cases and scoring formats, and more complex information on case design.

We have found that it is possible to put together a set of cases that might work in an examination setting. Clinicians were able to be trained in case writing (TB, SV) and wrote effective testing cases. The length of cases, between 5 and 15 minutes each, would allow 10–20 cases in all for a 2-hour exam, which would allow adequate sampling of the curriculum. In retrospect the cases appear too easy to discriminate adequately, as most students scored highly, and this would not allow discrimination between good and poor students. The complete absence of technical issues demonstrates how robust the application is, although scaling this up for 250 students might overload the vpSim server.

The focus groups demonstrated that the cases and format was perceived as fair, testing and appropriate by the students. They highlighted that clinical reasoning could be tested in this format, that linear and branched designs could test different skills, and that different stages in medical education required different formats.

This project has now allowed the group to develop more AVPs with a mixed linear/branching design for formative use. Scores will be correlated against the scores in end of year examinations to find out more about what the AVPs are actually testing. Branched AVPs are also being developed for incorporation into the final (exit) examination from the medical school.

At last perhaps a test will be devised that meet the characteristics of the ideal assessment (Table 1).

References

Andriole, D. A., Jeffe, D. B., Hageman, H. L., & Whelan, A. J. (2005). What predicts USMLE step3 performance? *Academic Medicine, 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. *British Medical Journal, 329*(7466), 606–609.

Conradi, E., & Round, J. (2008). *Virtual patients as a tool for assessment.* Medbiquitous Conference 2008. Retrieved March 3, 2010, from http://www.medbiq.org/events/conferences/2008/presentations.html

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. *Medical Teacher 30,* e66–e76.

Dent, J., & Harden, R. (2005). *A practical guide for medical teachers*. Churchill Livingstone.

Guagnano, M. T., Merlitti, D., Manigrasso, M. R., Pace-Palitti, V., & Sensi, S. (2002). New medical licensing examination using computer-based case simulations and standardized patients. *Academic Medicine, 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. *Medical Teacher, 30*(7), 673–678.

Huwendiek, S., Balasubramaniam, C., & Round, J. (2008). *Revip — An Anglo-German virtual patient case study exploring 'repurposing and enriching' as an effective way to share.* Prague: Association of Medical Education in Europe.

Kreiter, C. D., Ferguson, K., & Gruppen, L. D. (1999). Evaluating the usefulness of a computerized adaptive testing for medical in-course assessment. A*cademic Medicine, 74,* 1125–1128.

Sawhill, A. J., Dillon, G. F., & Ripkey, D. R. (2003). The impact of postgraduate training and timing on USMLE step 3 performance. *Academic Medicine, 78*(Suppl 1): S10–S12.