The death of the long case?

John J Norcini

The long case has tested the clinical skills of students for many years, yet inherent problems preclude it from being the sole source of decision making

American Board of Internal Medicine, Philadelphia, PA 19006, USA
John J Norcini
senior vice president
jnorcini@icemed.org

BMJ 2002;324:408–9

For most of the past century the long case has been an essential tool in the evaluation of clinical skills, being used to make important judgments about the education and practice of doctors throughout the world. Despite its widespread use and rich tradition, the long case has major flaws, making it unwise to use it as the sole basis for making decisions of consequence. In this article I describe the strengths and weaknesses of the long case and report on some of the modifications that have been introduced to improve its performance.

Summary points

The long case has been an essential tool in the assessment of clinical skills

It evaluates performance with real patients and enables students to gather information and develop treatment plans under realistic conditions

Problems have been found with the reproducibility of scores generated by the long case

The long case can be improved by increasing the number of encounters, examiners, or aspects of a competence assessed

Organisations responsible for high stakes testing are increasingly abandoning the long case or using it only in combination with other forms of assessment

The long case

Although there are numerous variations on the long case, traditionally a student is given unobserved time with a real patient in a clinical setting. During that time the student conducts an interview and performs a physical examination as appropriate. The student then presents his or her findings and plans to the examiners, who ask about the patient and related topics enabling them to judge the quality of the student's performance.

Strengths

The primary strength of the long case is that it evaluates the student's performance with a real patient. In the early stages of training, the objective structured clinical examination is often used, which assesses clinical skills well.[1] However, the standardised or simulated patients that are part of the examination are limited in the number and complexity of the medical problems they can portray. Consequently, as students approach entry to practice, assessment also needs to be based on performance with real patients who can exhibit the range of conditions seen in the clinical setting.

The long case also presents students with a complete and realistic clinical challenge. They are required to obtain all relevant information from the patient, structure the problem, synthesise their findings, and formulate a management plan. This contrasts with the typical objective structured clinical examination, with each station focusing on one aspect of the doctor-patient encounter.

Weaknesses

Implicit in the use of the long case is the assumption that if the student was examined again with another patient and different examiners, the results would be the same. Otherwise the scores could not be trusted to predict performance in practice, and it would make no sense to use them for assessment. The concept that test results should be able to be generalised or be repeatable is referred to as reliability or reproducibility.[2]

Over the past 30 years it has become increasingly apparent that the long case does not yield results that achieve reasonable levels of reproducibility. For example, in the early 1970s the American Board of Internal Medicine's oral examination for cardiovascular disease consisted of two long cases, each with two examiners.[3] Putting both cases together yielded a score that had a reproducibility coefficient of 0.39, meaning that 39% of the variability in scores was due to students' ability (signal) and 61% to errors of measurement (noise). When adjusted to prophesise what would happen with only one long case, the coefficient drops to 0.24, indicating that scores are composed of more than three times as much noise as signal. Other studies have obtained similar results, in stark contrast to clinical examinations such as the objective structured clinical examination, which often achieve reproducibility coefficients of 0.80 or better.[4 5] The table presents some typical findings for reproducibility of various formats.

Three major factors explain why the long case has problems with reproducibility.[8] In decreasing order of importance they are the case specificity of problem solving, differences between examiners, and variability in the aspects of an encounter evaluated.

Case specificity

For the scores from the long case to be reproducible, students must perform at the same level regardless of the patient they examine, yet physician performance varies from case to case. The case specificity of problem solving was identified by Elstein and colleagues.[9] It has been replicated in many studies.

These findings should not be surprising. Physicians know that they do not perform uniformly across all patient problems or even across different patients with the same problem. They have areas of relative strength and weakness, they respond differently to patients depending on their personal and professional experiences, and patients respond differently to them depending on a variety of factors. Therefore an assessment device must sample broadly across patients to generate scores that will generalise to typical performance.

Examiner effects

For scores to be reproducible, examiners must apply the same standards. Research shows that even experienced educators differ when assessing the same event.[10] This should not be surprising either. Physicians have legitimate and desirable differences in knowledge, standards, emphasis, and values. Likewise, they occasionally respond out of their own deficits of knowledge or the inappropriate influence of the patient's or student's style, appearance, race, sex, ethnicity, and so on. Further, patients' conditions naturally vary in difficulty, and the examiners must precisely compensate for the differences among them. These issues, and others, ensure that examiners differ when evaluating the same thing, undermining the reproducibility of the scores for the long case. An assessment device must sample across examiners to generate reproducible results.

Aspects of a competence assessed

For scores to be reproducible, it is important to assess several aspects of the competence being elicited by the student-patient encounter. Specifically, a variety of studies show that the information obtained from measurements is increased when examiners are instructed to evaluate a standardised list of different features of a competence, or when they observe the student-patient encounter rather than make a single global assessment or base their judgments on interrogation alone.[11–13] Again, these empirical results should not be surprising. Without specific instruction examiners will naturally attend to different aspects of an encounter, and this will be reflected in their evaluations. An assessment device must sample systematically across aspects of a competence to generate reproducible results.

Modifications to the long case

Without modification the traditional long case should not be used by itself to make critical decisions about the competence of a student. However, at least three strategies have been applied to improve its performance.

Encounters

Increasing the number of student-patient encounters is the single most important step in rehabilitating the long case. The addition of each encounter produces noticeable gains in the reproducibility of scores, even if there are limits on the number of examiners and aspects of competence involved. With enough encounters it is possible to achieve a high level of reproducibility and, without enough cases, all other changes will have only minimal impact. Where it is impractical to add long cases, one alternative is to add short cases, stations in the objective structured clinical examination, or other forms of assessment to broaden the sample of student-patient encounters.

Reproducibility of assessment formats studied by the American Board of Internal Medicine (estimates, based on three hours' testing time, will vary in other settings depending on quality of test material and heterogeneity of examinees)

Format	No of cases or items	Reproducibility coefficient
Oral examination (long cases)	2	0.39
Computer simulation (long cases)*	3	0.55
Written simulation (long cases)*	6	0.70
Miniclinical evaluation exercise (short cases)†	9	0.73
Multiple choice questions: single best answer*	90	0.88

*Based on data found in studies by Norcini et al.[6]

†Based on data found in studies by Norcini et al.[7]

Examiners

Examiners can be manipulated in at least three ways to improve the reproducibility of scores: by employing a statistical model that removes differences among them, by training them, or by increasing their numbers. By themselves these strategies produce only modest gains in reproducibility. Of them, increasing the number of examiners will have the largest effect, although gains will be minimal beyond a certain point (say four or five examiners).

Aspects of a competence

Increasing the number of aspects of a competence assessed and standardising them across examiners has a modest positive effect on the reproducibility of scores. For the long case this has included having the examiners observe the student-patient interaction or providing them with a list of competencies to evaluate. Again, the impact of this will be minimal beyond a certain point (say 5 or 10 aspects of a competence).

Competing interests: None declared.

1 Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ* 1979;13:41-54.

2 Crocker L, Algina J. *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston, 1986.

3 Meskauskas JA. Studies of the oral examination: the examinations of the subspeciality Board of Cardiovascular Disease of the American Board of Internal Medicine. In: Lloyd JS, Langsley DG, eds. *Evaluating the skills of medical specialists*. Chicago, Il: American Board of Medical Specialties, 1983.

4 Waas V, Jones R, van der Vleutin C. Standardized or real patients to test clinical competence? The long case revisited. *Med Educ* 2001;35:317-8.

5 Yang JC, Laube MD. Improvement of reliability of an oral examination by a structured evaluation instrument. *J Med Educ* 1983;58:864-72.

6 Norcini JJ, Meskauskas JA, Langdon LO, Webster GD. An evaluation of a computer simulation in the assessment of physician competence. *Eval Health Prof* 1986;9:286-304.

7 Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Intern Med* 1995;123:795-9.

8 Norcini JJ. The validity of long cases. *Med Educ* 2001;35:735-6.

9 Elstein AS, Shulman LS, Sprafka SA. *Medical problem-solving: an analysis of clinical reasoning*. Cambridge, MA: Harvard University Press, 1978.

10 Noel GL, Herbers JE, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents? *Ann Intern Med* 1992;117:757-65.

11 Waas V, Jolly B. Does observation add to the validity of the long case? *Med Educ* 2001;35:729-34.

12 McKinley RK, Fraser RC, van der Vleuten C, Hastings AM. Formative assessment of the consultation performance of medical students in the setting of general practice using a modified version of the Leicester Assessment Package. *Med Educ* 2000;34:573-9.

13 Kroboth FJ, Hanusa BH, Parker S, Coulehan JL, Kapoor WN, Brown FH, et al. The inter-rater reliability and internal consistency of a clinical evaluation exercise. *J Gen Intern Med* 1990;5:214-7.

(Accepted 23 January 2002)