The long case and its modifications: a literature review

Gominda G Ponnamperuma,[1] Indika M Karunathilake,[2] Sean McAleer[1] & Margery H Davis[1]

CONTEXT This review provides a summary of the published literature on the suitability of the long case and its modifications for high-stakes assessment.

METHODS Databases related to medicine were searched for articles published from 2000 to 2008, using the keywords 'long case', 'clinical examinations' and 'clinical assessment'. Reference lists of review articles were hand-searched. Articles related to the objective structured clinical examination were eliminated. Research-based articles with hard data were given more emphasis in this review than those based on opinion.

RESULTS Eighteen articles were identified. The main disadvantage of the long case is its inability to sample the curriculum widely, resulting in low reliability. The main advantage of the long case is its ability to assess the candidate's overall (holistic) approach to the patient. Modifications to the long case attempt to: structure the format and the marking scheme; increase the number of examiners; observe the candidate's behaviour, and increase the number of cases.

CONCLUSIONS The long case is a traditional clinical examination format for the assessment of clinical competence and assessment at this level is important. The starting point for the majority of recent research on the long case has been an acceptance of its low reliability and modifications to the format have been proposed. Further evidence of the efficacy of these modifications is required, however, before they can be recommended for summative assessment. If further research is to be undertaken on the long case, it should focus on finding practicable ways of sampling the curriculum widely to increase reliability while maintaining the holistic approach towards the patient, which represents the attraction of the long case.

*Medical Education 2009: **43**: 936–941*
doi:10.1111/j.1365-2923.2009.03448.x

[1]Centre for Medical Education, University of Dundee, Dundee, UK
[2]Medical Education Development and Research Centre, Faculty of Medicine, University of Colombo, Colombo, Sri Lanka

Correspondence: Gominda G Ponnamperuma, 16/5 Quarry Road, Mirihana, Nugegoda, Sri Lanka. Tel: 00 94 1128 27531; Fax: 00 94 1126 91581; E-mail: gominda@googlemail.com

INTRODUCTION

The long case enjoys a unique place in many clinical assessment systems and continues to be used in both undergraduate and postgraduate medical education in many parts of the world. The 2009 Australian publication *Mastering the Medical Long Case* by Jayasinghe[1] was recommended on the basis that 'all medical schools have recognised the ''Long Case'' as an integral part of the learning and examination process of the medical program'. Use of the long case in high-stakes assessment continues, despite concerns raised about its validity and reliability. The continued popularity of the long case stimulated us to review the literature to investigate its pros and cons. This review also looks into possible alternatives to the long case, as well as ways and means of improving it.

The long case is a traditional clinical examination that assesses candidate competence at the 'shows how' level in Miller's pyramid.[2] The candidate initially spends time (30–60 minutes) with a patient, taking a history and carrying out physical examination, without examiner observation. Then the candidate presents the findings to one or more examiners and answers oral questions. In most instances each candidate is given a unique patient and a unique examination. Traditionally, the candidate is scored with unstructured marking criteria that are based on neither standardised checklists nor on rating scales with descriptors related to candidate competence.

Prior to the turn of this century, two problems were identified in relation to the long case. Firstly, Wilson *et al.*[3] found some substantial differences in scores given to the same candidate by different examiners in an undergraduate clinical examination in Glasgow, UK, resulting in low validity and reliability. Secondly, van der Vleuten[4] and Dugdale[5] pointed out that the long case attempted to generalise the abilities of the candidate across a broad spectrum of clinical scenarios with a single clinical case. This problem has been confirmed by more recent studies.[6,7]

Given the above two problems, the main research questions investigated through this literature review were: What modifications to the long case have been attempted with a view to improving its psychometric properties? What are the advantages and disadvantages of the long case?

METHODS

MEDLINE, BIDS, Blackwell Synergy (for the journals *Medical Education* and *Internal Medicine*), Ingenta, PubMed, AskEric, TimeLit and Google Scholar were searched for articles published from 2000 to 2008. The keywords were 'long case', 'clinical assessment' and 'clinical examinations'. The reference lists of review articles were then hand-searched. This hand-search identified articles published prior to 2000. Those that directly dealt with the long case and related clinical assessments were selected for this review.

RESULTS

Although it is much used, there is little published research on the long case. Eighteen articles directly related to the 'traditional' long case were found. The findings of these articles and the pre-2000 papers identified by the hand-search are discussed under three headings: modifications to, advantages of, and disadvantages of the long case.

MODIFICATIONS TO THE LONG CASE

The observed structured long case

Gleeson[8,9] introduced the objective structured long examination record (OSLER), a 10-item analytical record of the traditional long case, with an examiner-observed history-taking and physical examination process, and a criterion-referenced marking scheme to improve the reliability of the long case. No reliability figures for the OSLER were reported. In postgraduate clinical skills assessment, Gleeson[8] reported the ability of the OSLER to identify the curriculum content that needed more input by the curriculum designers. Van der Vleuten[4] noted that the OSLER was strong in educational value in terms of providing feedback. He pointed out, however, that improvements in reliability were better achieved by increasing the number of cases than by focusing on observing the student during the long case.

In a study of doctor trainees with an observed long case and a structured assessment form, Pavlakis and Laurent[10] found that the trainees did not pay attention to physical examination skills as these skills were not previously observed. The study upheld the value of observation of the long case as it compelled the candidates to master clinical skills. The authors

were critical of the importance placed on the discussion of patient management in the long case at the expense of the assessment of clinical examination technique.

The structured long case with multiple examiners

Olson *et al.*[11] evaluated the usefulness of a structured question grid for the long case, where two assessors assessed the candidate on one long case. One examiner marked with the aid of a structured question grid and the other did not. Based on the results of 391 students taking 1564 long cases in internal medicine, paediatrics, reproductive medicine or surgery, there was no significant difference 'in the chance of students being assessed as failing or on the likelihood of a discrepancy between the ratings'. The student group that was assessed with the structured question grid, however, perceived the assessment as less representative of their ability.

The observed structured long case with multiple examiners

Wass and Jolly[6] experimented with the traditional history-taking long case by using two pairs of examiners and incorporating examiner observation into the final MBBS examination at a London medical school. A pair of examiners first observed the candidate taking the history (Part 1). Thereafter, the candidate presented the case to another pair of examiners (Part 2). Inter-rater reliability was higher for the observation (checklist 0.72; global 0.71) than the presentation (checklist 0.38; global 0.60) part. The authors also found that observation of the long case history taking constituted a distinct component of clinical competence, which the usual long case (i.e. only the presentation part) would fail to measure.

Norcini,[12] however, argued that although experiments similar to that conducted by Wass and Jolly[6] improved reliability, these modifications did not raise the long case to a level that supported its use in high-stake situations. Three factors that contributed to its unreliability were: case specificity; examiner stringency, and the aspects of competence evaluated.[13] Norcini proposed that the modification of these three factors, respectively, would bring about substantial gains in reliability, as follows:

1 cases or encounters: by increasing the number of cases or encounters;
2 examiners: by minimising differences among examiners; by increasing the number of examiners, and by training the examiners, and
3 aspects of competence: by increasing the number of aspects of competence assessed and providing the examiners with lists of competencies; by standardising across examiners through examiner training, and by using examiner-observed student–patient interactions.

Price and Byrne[14] assessed clinical psychiatry skills, where two examiners first directly observed the candidate taking a history for 20 minutes and then evaluated candidate competence on case-specific tasks, such as partial mental state assessment, as requested by the examiners. The key feature of this method was that it allowed the examiner to adjust for the degree of difficulty of the case. Although the authors reported satisfactory inter-rater reliability (kappa coefficient of 0.7), this study did not address the low generalisability problem of the long case. The study only partially addressed the problem of lack of standardisation by allowing the examiners to adjust the scoring according to the level of difficulty of the case.

Increasing the number of cases: multiple observed structured long cases with multiple examiners

Improvements to the observed long case include the direct observation clinical encounter examination (DOCEE)[15] and the integrated direct observation clinical encounter examination (IDOCEE).[16] Both examinations expose the candidate to multiple patient interactions in which multiple examiners from different specialties observe the candidate carrying out history taking and physical examination. In the DOCEE, each candidate was examined with four patients and two pairs of examiners. Each pair of examiners assessed the candidate in two patient encounters. Every three consecutive candidates were examined with the same set of patients and examiners. The generalisability coefficient for four cases and two examiners for each case was 0.84.[15] The IDOCEE was very similar in structure and conduct to the DOCEE, except for the number of patients (four to six) and examiners (two panels, each with two or three examiners) encountered by each candidate. Each panel of examiners assessed a candidate in two or three patient encounters. The students and examiners were 'highly satisfied' with the structure, organisation and effectiveness of this examination.[16]

In a separate experiment with two observed long cases and a pair of examiners, Newble[17] demonstrated the effectiveness, as measured by student and staff feedback. Luiz *et al.*[18] also found that when each candidate took two structured, standardised,

observed long cases, each marked by a different examiner, examiner agreement on candidate achievement of clinical skills was 89%. Wass *et al.*[19] experimented with two observed long cases, each examined by two examiners. They found that if each long case was observed by one unique examiner, at least 10 observed history-taking long cases were required to achieve 0.8 reliability. Each long case lasted 21 minutes (a 14-minute patient encounter followed by a 7-minute interview) and each candidate encountered eight different examiners and patients. It should be stressed, however, that this modification applied to only the history-taking component of the conventional long case.

ADVANTAGES OF THE LONG CASE

Authenticity

The long case provides an interaction between the candidate and the patient[6] that integrates history taking, physical examination, investigation, diagnosis and management. The candidate needs to obtain relevant information, structure a problem, synthesise the findings and formulate a management plan.[13]

Because 'real' patients are used, the long case is more authentic (i.e. it represents a real-life experience) than simulated patient scenarios can be[4,6,13] and hence has greater validity[5,6,20] in that it provides a real-time, actual patient problem, which must be approached holistically. Furthermore, it offers direct contact between the candidate and the examiner.[4]

Educational value

The long case is an educationally valuable test[4,6] because it provides diagnostic feedback to both students and teachers. The long case is a good method of formative assessment.[4,12] Teachers can use the results of the long case to identify any neglected areas or teaching deficiencies based on course outcomes.[10] The long case can also be used as a screening device to identify weak students for remediation.[12,17] It has proven to be useful in evaluating the effectiveness of educational programmes.[8–10,12]

DISADVANTAGES OF THE LONG CASE

Non-generalisability

A good result in one long case does not guarantee a similarly high result in another long case. The result of one long case is not a generalisable indicator of the candidate's ability across a range of other cases and clinical situations.[4–6] Van der Vleuten,[4] summarising its inappropriateness, reiterated: 'We intuitively believe that when we have measured someone's competence with one patient we can predict how competent that person will be with another. Unfortunately, this prediction tends to be poor, and it is this factor that leads to serious unreliability.' Dugdale[5] then drove the point home, saying: '…if a doctor failed to diagnose my (hypothetical) prostatic carcinoma, it would be small consolation to know that he had done brilliantly in his clinical long case on multiple sclerosis.' The inability to assess candidate competence through a single case has been termed 'case specificity'. Many authors[4,6,16] have emphasised that one long case does not offer a sufficiently representative sample of cases to measure examinee competence.

Olson,[21] however, observed that, except for borderline students, a single long case in one discipline was good enough to predict performance in internal medicine, paediatrics, reproductive medicine or surgery. He arrived at this conclusion by comparing long case marks in four disciplines, obtained over 6 years. He also found that for borderline students, two cases would be sufficient to predict the outcome in the four disciplines. This is the only study we found supporting the generalisability of a single long case. This finding must be balanced against a large body of evidence and opinions[4–6] to the contrary.

Low in reliability

Van der Vleuten,[4] Gleeson,[9] Olson *et al.*,[11] Norcini,[12,13] Abouna and Hamdy[16] and Paul[22] all identified the poor reliability of the results of the long case. A recent study[23] in postgraduate assessment estimated that at least five or six, 85-minute (a candidate spends 60 minutes with the patient and 25 minutes with the examiners) long cases were necessary to achieve 0.8 dependability, which is a more conservative figure of reliability (i.e. dependability takes into account the variance contributed to the measurement error by factors or facets not directly associated with the candidate, such as cases, examiners, and interaction between cases and examiners). This study also found that, with two such long cases, the percentage variability (i.e. variance) explained by the candidate's ability and the interaction between cases and candidates were similar. The latter finding confirms that case specificity[24] is a major contributor to the (un)reliability of the long case. 'Over the past 30 years,' suggested Norcini,[13] 'it has

become increasingly apparent that the long case does not yield results that achieve reasonable levels of reproducibility.'

Norman[25] commented that the long case may have slightly better reliability than the objective structured clinical examination (OSCE) if it is conducted as observed, multiple long cases. As Wass *et al.*[19] found, the reliability of the long case, when carried out with two pairs of examiners, was not better or worse than that of the OSCE. They estimated that a reliability of 0.8 can be achieved with 10 long cases on history taking with two examiners observing each long case. However, the time, logistics and cost-effectiveness issues related to running multiple long cases with multiple examiners preclude their use in standard examinations.

Low in validity

The low validity of the long case stems from its inability to generalise from the results of a single patient interaction, non-observation of the candidate during the patient interaction, lack of structure, and lack of patient standardisation.

Attempting to judge competence across a range of clinical conditions on the basis of one unobserved case[9] is a major contributor to low content validity. Although Olson's[21] findings suggested that this was possible except for borderline candidates, this is a highly important exception as it is the borderline candidate who presents the greatest assessment challenge.

As the traditional long case is not observed by the examiner, it assesses history-taking ability, communication skills[6,16,26] and physical examination skills only poorly.[10]

The low validity of the traditional long case may also be related to a lack of structure[6] that leads to global pass or fail decisions.[26] As identified above, however, there have been various moves to structure the long case.

The lack of patient standardisation and heavy dependence on the 'luck of the draw'[9,26] (different candidates are assessed on different patients) may contribute to the low validity of the long case. As the long case did not allow examiners to adjust according to the degree of difficulty posed by the patient, Price and Byrne[14] described a modification, in which the examiners first marked patient difficulty on a 10-point rating scale independently, and then marked the candidate performance on a 7-point rating scale, initially independently, followed by consensus. Unfortunately, the authors did not reveal how the difficulty rating was used to modify the candidates' scores. The candidates found the examination stressful, but rated the method as appropriate for clinical assessment.

Feasibility, efficiency and cost-effectiveness

Although it is a lengthy examination, Wass and Jolly[6] indicated that the time taken to assess a candidate in the long case may be an advantage. However, most examination time is not spent on examining the candidate, but represents unobserved time that the candidate spends with the patient.[13] The above authors were doubtful whether this long time period is put to good use in terms of assessment of outcomes or agreed competencies. Abouna and Hamdy[16] observed that their version of the long case, the IDOCEE, was more cost-effective than the traditional long case.

In summary, the modifications to, and the advantages and disadvantages of the long case indicate that the main impediment to improving the validity and reliability of the long case concerns the overly long time the examination takes, which poses challenges to sampling the assessment content more broadly (i.e. by introducing more cases).

CONCLUSIONS

Various modifications to increase the number of examiners and cases and to standardise and structure the long case have been attempted. More evidence of the effectiveness of these modifications is required, however, before they can be recommended for summative assessment.

The advantages of the long case include its authenticity in assessing candidate competence holistically, and its educational use in providing feedback to the candidate, teacher and institution about the curriculum, teaching and candidate ability. Its disadvantages include the inability to generalise from one long case about the candidate's ability in other cases, poor reliability, low validity, and the long duration of the examination. These disadvantages preclude its use in high-stakes examinations unless and until a more psychometrically and educationally desirable format of the long case has been devised to accommodate multiple, observed patient interactions with multiple examiners.

If sampling of more cases can be incorporated into the long case format, then the main advantage of the long case can be utilised (i.e. it is a realistic, holistic and in-depth assessment of patient encounter). Although the possibility of introducing more cases to the long case format might be explored in future research, the value of such studies is debatable when set against the practicability of an examination comprising multiple long cases.

Contributors: GGP performed the literature review, wrote the first draft of the paper and made subsequent amendments to the original manuscript. IMK contributed to the literature search and commented critically on the draft manuscript. SM and MHD contributed to the draft manuscript, read and critically commented on the subsequent manuscript, and recommended revisions. All authors approved the final manuscript for publication.
Acknowledgements: none.
Funding: this study was internally funded.
Conflicts of interest: none.
Ethical approval: not required.

REFERENCES

1 Jayasinghe R. *Mastering the Medical Long Case*. Chatswood, New South Wales: Churchill Livingstone 2009.
2 Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;**65** (9) (Suppl):63–7.
3 Wilson GM, Lever R, Harden RM, Robertson JIS, MacRitchie J. Examination of clinical examiners. *Lancet* 1969;**1** (7584):37–40.
4 van der Vleuten CPM. Making the best of the 'long case'. *Lancet* 1996;**347** (3):704–5.
5 Dugdale A. Letters: long case clinical examinations. *Lancet* 1996;**347**:1335.
6 Wass V, Jolly B. Research paper: does observation add to the validity of the long case? *Med Educ* 2001;**35**:729–34.
7 Wass V, van der Vleuten C. The long case. *Med Educ* 2004;**38**:1176–80.
8 Gleeson F. Defects in postgraduate clinical skills as revealed by the objective structured long examination record (OSLER). *Ir Med J* 1992;**85** (1):11–4.
9 Gleeson F. AMEE Medical Education Guide No. 9: assessment of clinical competence using the objective structured long examination record (OSLER). *Med Teach* 1997;**19** (1):7–14.
10 Pavlakis N, Laurent R. Role of the observed long case in postgraduate medical training. *Intern Med J* 2001;**31** (9):523–8.
11 Olson LG, Coughlan J, Rolfe I, Hensley MJ. The effect of a structured question grid on the validity and perceived fairness of a medical long case assessment. *Med Educ* 2000;**34** (1):46–52.
12 Norcini J. The validity of long cases. *Med Educ* 2001;**35**:720–1.
13 Norcini JJ. The death of the long case? *BMJ* 2002;**324**:408–9.
14 Price J, Byrne JA. The direct clinical examination: an alternative method for the assessment of clinical psychiatry skills in undergraduate medical students. *Med Educ* 1994;**28** (2):120–5.
15 Hamdy H, Prasad K, Williams R, Salih FA. Reliability and validity of the direct observation clinical encounter examination (DOCEE). *Med Educ* 2003;**37** (3):205–12.
16 Abouna GM, Hamdy H. The integrated direct observation clinical encounter examination (IDOCEE) – an objective assessment of students' clinical competence in a problem-based learning curriculum. *Med Teach* 1999;**21** (1):67–72.
17 Newble DI. The observed long case in clinical assessment. *Med Educ* 1991;**25** (5):369–73.
18 Luiz EAT, Roberto OD, Fernando CF, Eduardo F, Lio CM, Ana LCM, Lio CV. A standardised, structured long case examination of clinical competence of senior medical students. *Med Teach* 2000;**22** (4):380–5.
19 Wass V, Jones R, van der Vleuten CPM. Standardised or real patients to test clinical competence? The long case revisited. *Med Educ* 2001;**35**:321–5.
20 Smee S. ABC of learning and teaching in medicine: skill-based assessment. *BMJ* 2003;**326**:703–6.
21 Olson LG. The ability of a long case assessment in one discipline to predict students' performances on long case assessments in other disciplines. *Acad Med* 1999;**74** (7):835–9.
22 Paul VK. Assessment of clinical competence of undergraduate medical students. *Indian J Pediatr* 1994;**61** (2):145–51.
23 Wilkinson TJ, Campbell PJ, Judd SJ. Reliability of the long case. *Med Educ* 2008;**42**:887–93.
24 Eva KW. On the generality of specificity. *Med Educ* 2003;**37** (7):587–8.
25 Norman G. Editorial: the long case versus objective structured clinical examinations. *BMJ* 2002;**324**:748–9.
26 Sood R. Long case examination – can it be improved? *J Indian Acad Clin Med* 2001;**2** (4):251–5.

Received 23 December 2008; editorial comments to authors 25 February 2009; accepted for publication 5 June 2009