Education in Medicine Journal
2012, VOL 4 ISSUE 1
DOI:10.5959/eimj.v4i1.9

REVIEW ARTICLE

A literature review of the long case and its variants as a method of assessment

Simon Thornton

Clinical Teaching Fellow, North Bristol NHS Trust and University of Bristol, United Kingdom

Abstract

Objective: The purpose of this literature review is to review the evidence base for the long case and its variants as a reliable method of assessing medical students.

Method: MEDLINE, PubMed, EMBASE, Blackwell Synergy, Ask Eric and Google Scholar were searched for articles published between 1991 and 2011 using the keywords 'long case', 'OSLER', 'clinical assessment' and 'clinical examinations'.

Result: 29 articles were identified, 13 original research articles and 16 review articles and commentaries. The majority of studies are single centre cohort studies, providing relatively weak evidence. The strengths and weaknesses of the long case are discussed with a review of the literature.

Conclusion: Despite it being viewed as a relatively authentic examination, concerns regarding reliability of the long case examination has led to its use being discontinued in many UK medical schools. However, there is evidence that with simple modifications, the reliability of the long case can be improved to be as good as, if not better than alternative clinical examinations such as the Objective Structured Clinical Examination (OSCE).

Keywords:
Long case, OSLER, Assessment, DOCEE.

How to cite this article?

Thornton, S. (2012). A literature review of the long case and its variants as a method of assessment. *Education In Medicine Journal*, *4*(1). DOI:10.5959/eimj.v4i1.9

Education in Medicine Journal
2012, VOL 4 ISSUE 1
DOI:10.5959/eimj.v4i1.9

Introduction

The long case as a clinical examination emerged in Cambridge in the mid-19th Century as an assessment tool for clinicians at the university [1]. In its original form, it consists of around one hour of a candidate taking an unobserved history and examination of a real patient before presenting to an examiner in an unstructured manner [2]. The current final year long case assessment at our institution involves three long cases spread over four months. The first two are observed by a single examiner and are worth 15% each. The third long case is worth 70% and is assessed by two examiners.

The examination is criterion referenced, with students having to have an average mark of >50% to pass. It has often proved a contentious examination amongst both students and educators due to its perceived poor reliability [2, 3]. For this reason it has largely been discontinued in North America, though continues to be used in the UK and Australasia. It is often viewed as very much a 'luck of the draw' exam – if you get a 'good' patient and examiner, you are much more likely to pass. This is something that hasn't gone unnoticed by a number of commentators [4, 5]. A literature review was conducted to identify the evidence base for the long case as well as potential improvements.

Method

MEDLINE, PubMed, EMBASE, Blackwell Synergy, Ask Eric and Google Scholar were searched for articles published between 1991 and 2011 using keywords 'long case', 'OSLER', 'clinical assessment' and 'clinical examinations'. Articles relating to the long case were identified and included in the review.

Result

29 articles were identified. 13 of these were original research articles. The other 16 consisted of review articles and commentaries. A summary of the original research articles is provided in Appendix 1. The majority of studies are single centre cohort studies, providing relatively weak evidence.

A discussion of the characteristics of an ideal clinical assessment is given, followed by the findings of these articles which are discussed under the headings of variations to, strengths, and weaknesses of the long case.

Characteristics of an ideal clinical assessment

Before examining the evidence base behind the long case and its variants, it's important to have an understanding of what makes a good assessment. Reliability and validity are often used as quality indicators in educational assessment, drawing an analogue with sensitivity and specificity in clinical research.

Reliability refers to the ability of a test to consistently measure what it's supposed to measure [6]. When assessing the long case, we can subdivide this into consistency, test-retest reliability (or as it's often referred to in the long case: inter-case reliability) and inter-rater reliability (do different examiners give the same or different marks?).

Validity is how well the test measures what it's supposed to measure [6]. Ideally, a clinical assessment should sample the curriculum in a representative way (content validity, or sometime referred to as 'blueprinting'), make it clear what is asked of the students (construct validity), and is often expected to drive learning (consequential validity).

Unfortunately no clinical assessment method has been devised that is able to measure all facets of clinical competence [7] .
Miller's pyramid of competence highlights some of the issues involved with analysing validity [8], with the long case designed to assess the 'shows how' level. In an attempt to improve the reliability and validity of the long case, a number of variants have been tried.

Variations of the long case

The structured long case

The Objective Structures Long Examination Record (OSLER) was introduced by Gleeson [9] as a method to try and introduce better standardisation to the long case. The student conducts an hour long observed history and examination with a patient followed by 20-30 minutes of structured questioning by the examiner using a 10 item analytical record. As part of an effort to reduce the 'luck of the draw' aspect, examiners are asked to formally document the difficulty of the case. Unfortunately there is no evidence as to the reliability and validity of the OSLER, and to quote Norman;

'*Current discussions about best evidence medical education are an indication that, just as in clinical medicine, intuitions will frequently be at variance with evidence'.*[10]

Other methods have been devised that attempt to provide an evidence base for various forms of structured marking schemes. Luiz *et al* assessed 27 final year undergraduates on 2 long cases consisting of 25 minutes observed history and examination with 8 minutes of questioning, using a standardised 10 item checklist. With each case having a different examiner, they found an 89% rate of agreement between different examiners on assessment of various skills of the same patient [11].

Other attempts to introduce more standardised criteria have been less successful. Olson *et al* developed a 'Structured Question Grid' whereby examiners assessed the patients in advance and wrote down points they wanted the student to specifically pick up on [12].

The history and examination consisted of an hour of unobserved history and examination, followed by 30 minutes of questioning and presentation. 67 students were randomised to either the 'standard' long case format, or the 'Structured Question Grid'. There was no difference in failure rate, however students whose assessors used the grid felt constrained by the use of a formal mark scheme.

The observed long case

Many long case variants now have an observed history and examination component [9, 11, 13]. Wass and Jolly studied a modified version of the long case – the history taking (HT) long case [13]. This consisted of 16 minutes of observed history taking by one or two examiners, followed by an eight minute presentation to one or two different examiners. They found that inter-rater reliability was higher for observation than for presentation, suggesting observation of the long case is a more useful component than the presentation.

The long case with multiple examiners

It would seem logical that having more examiners would reduce inter-rater variation, and the fears that students often have of having to pander to the nuances of their particular examiner. It would also reduce the chance of an examiner responding out of inappropriate influence of the student's sex or ethnicity [5]. A study by Wilkinson *et al* assessed postgraduate students in two unobserved long cases of an hour, followed by 25 minutes with two examiners [14].

Different examiners were used for the two different long cases. An in depth statistical analysis showed that only 10% of result variation was due to problems with inter-rater reliability. Furthermore, Hamdy *et al* produced a staggering figure using generalisability theory showing that using their particular protocol, 10 examiners would produce a generalisability coefficient of 0.61 and that increasing the number of examiners to 50 would only produce an increase to 0.62 [15]. These coefficients should ideally be >0.8.

Multiple long cases, multiple examiners

Perhaps the best example of this is provided by Hamdy *et al* in their Direct Observation Clinical Encounter Examination (DOCEE). In a study of 56 final year undergraduates in Bahrain, they introduced a complex exam that had a number of important modifications [15]. The history taking and examination length was shortened to 30 minutes with subsequent questioning for 15 minutes by two examiners, one specialist and one non-specialist. Four long cases were done in total, from four different, well defined areas of medicine, surgery, paediatrics and women's health. Each panel of two examiners assessed two of the long cases for the same student.

The student was then evaluated using a structured checklist. Every third consecutive candidate was examined with the same set of patients and examiners. This produced a generalisability coefficient of 0.84 for 4 observed long cases with 2 examiners. This compares very favourably with the reported 0.36 for a single long case with single examiner [16] .

Strengths of the long case

Authenticity

One of the key perceived advantages of the long case is its authenticity [2-5, 14]. The long case presents a realistic challenge and assesses the student's overall ability to carry out a medical history, examine and communicate with a patient. It's depth allows the examiner to assess all domains of knowledge, skills and attitudes [6]. However, a one hour long case as exists at our institution is less authentic. Once students start work, they have to learn very quickly to work efficiently and an hour is arguably too long for a history and examination. Further on in general practice training, some may have only 10 minutes. It's also limited to a 'new presentation' scenario. Most clinic appointments are for follow up – not new patients.

Consequential validity

Perhaps the biggest advantage of the long case is in its educational value. Despite the best efforts of curriculum designers to develop a broad syllabus, it is the examinations that drive learning to a great extent [17, 18]. If this holds true, the long case is absolutely crucial to medical education as there is no better way of preparing for practice than repeatedly taking a history from, and examining real patients. As part of practicing for the real assessment, we often observe 'mock' long cases with the students.

This provides a unique opportunity for the student to be observed one-to-one with a clinician and receive extensive feedback on their performance. Students often report that this is one of the most valuable educational sessions they have in their final year at medical school – despite what their views may be of the long case as a method of assessment.

Weaknesses of the long case

Low reliability

Many authors have identified the low reliability of the long case as a method of clinical examination [5, 12, 17]. There are serious concerns about the ability of a candidate in one long case to perform well in another. As Dugdale says,

'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.' [18]

A number of studies have tried to address this issue [14, 19]. Returning to Wilkinson's study of postgraduate students in Australia consisting of two long cases per student with two different examiners, although only 10% of result variation was due to inter-rater reliability, 37% was due to inter-case (or test-retest) reliability. Only 38% of result variation was due to candidate ability [14]. It was estimated that five or six unobserved cases would be needed to achieve a dependability of 0.8.

A number of other studies have noted that inter-rater reliability is a minor problem in relation to inter-case reliability [5, 11, 20]. Interestingly however, in a study by Olson of 391 undergraduate students in Australia, a passing grade on a single long case in any discipline reliably identified a majority of students as having a <10% risk of poor performance [19], and by increasing the number of observed cases to 4, a high generalisability coefficient of 0.84 has been demonstrated [15].

Low validity

One of the main problems with the long case, which is intertwined with inter-case reliability, is its content validity. In a single long case, it is impossible to sample the whole curriculum. There is the real danger of having a patient with the weird and wonderful on the fringes of the curriculum. This has been overcome to a degree by Hamdy *et al* with their DOCEE. The four patients they select for the exam are from a pre-prepared blueprinted list of common health problems/presentations and cover four areas: women's health, paediatrics, surgery and medicine [15]. The use of this assessment specification allows greater sampling of the curriculum and abilities to be tested [6].

Practicability

The main difficulty with organising long case examinations is finding enough patients for the number of students. It is fairly cost effective compared to an OSCE as the patient is essentially a 'free' resource, compared to the cost of hiring actors and buying disposable equipment for the OSCE.

Alternatives to the long case

The Objective Structured Clinical Examination (OSCE)

The OSCE uses short, standardised stations focusing on specific skills and often substituting real patients with standardised simulations [21]. The move towards OSCEs has been driven by the perceived low reliability of the long case. Studies have however demonstrated that, if the time given to both examinations were equal, the long case would be just as reliable, if not more so, than the OSCE (0.84 reliability for 3.5 hours of long case vs 0.73 for 3.5 hours of OSCE [21]). Thus the practicability of the long case is similar to the OSCE. Furthermore the OSCE lacks authenticity compared to the long case, and it lacks the complex interplay that defines the doctor-patient relationship:

'Could we conceive of a professional music student who is told that her final acceptability as a musician will depend on a series of assessments of scales and short pieces but never a recital of a complete piece of music? [22]'

Mini-Clinical Evaluation Exercise (CEX)

This provides a snapshot evaluation of student or doctor/patient interaction. It is often used for formative rather than summative assessment in postgraduate training as it needs to be repeated frequently to achieve reliability [23].

Does the long case have a future?

This review article was conducted with a view to potentially improving the long case as a method of examination at our institution. The evidence suggests that this could be achieved by:

1. Use of a more formal structured checklist to assess several measures of clinical competence.
2. Formative assessment of several long cases over a longer time period.
3. The introduction of a shorter time for focused history and examination to improve authenticity and allow for a greater number of cases to be seen (point 4).
4. Increase the number of long cases seen by each student to around 4, similar to Hamdy *et al*, to dramatically improve reliability.
5. Communicate the evidence base for the long case to students to allay some of the common misconceptions surrounding it.

Reference

1. A. Dare, A. Cardinal, J. Kolbe and W. Bagg, What can history tell us? An argument for observed history-taking in the trainee intern long case assessment., *The New Zealand Medical Journal* **121** (2008).
2. S.Smee, ABC of learning and teaching in medicine. Skill based assessment., *British Medical Journal* **326** (2003).
3. N. Teoh and F. Bowden, The case for resurrecting the long case, *British Medical Journal* **336** (2008).
4. R. Sood, Long Case Examination - Can it be Improved?, *Journal of the Indian Academy of Clinical Medicine* **2** (2001).
5. J. Norcini, The death of the long case?, *British Medical Journal* **324** (2002). 6. I. Reece and S. Walker, Teaching, Training and Learning, Business Education Publishers, Sunderland (2007).
7. V. Wass, C. Van der Vleuten, J. Shatzer and R. Jones, Assessment of clinical competence, *Medical Education Quartet* **357** (2001), pp. 945-949.
8. G. Miller, The assessment of clinical skills/competence/performance., *Academic Medicine* **65** (1990).
9. F. Gleeson, Assessment of clinical competence using the Objective Structured Long Examination Record (OSLER), *Medical Teacher* **19** (1997), pp. 7 14.
10. G. Norman, The long case versus objective structured clinical examinations, *British Medical Journal* **324** (2002), pp. 748-749.
11. T. Luiz, R. Dantas, J. Fernando, E. Ferriolli, J. Moriguti, A. Martinelli, *et al.*, A standardized, structured long-case examination of clinical competence of senior medical students., *Medical Teacher* **22** (2000), pp. 380-385.
12.L. Olson, J. Coughlan, I. Rolfe and M. Hensley, The effect of a Structured Question Grid on the validity and perceived fairness of a medical long case assessment, *Medical Education* **34** (2000), pp. 46-52.

13.V. Wass and B. Jolly, Does observation add to the validity of the long case?, *Medical Education* **35** (2001), pp. 729-734.
14.T. Wilkinson, P. Campbell and S. Judd, Reliability of the long case, *Medical Education* **42** (2008), pp. 887-893.
15.H. Hamdy, K. Prasad, R. Williams and F. Salih, Reliability and validity of the direct observation clinical encounter examination (DOCEE), *Medical Education* **37** (2003), pp. 205-212.
16.T. Wilkinson, L. D'Orsogna, B. Nair, S. Judd and C. Frampton, The reliability of long and short cases undertaken as practice for a summative examination, *Internal Medicine Journal* **40** (2010), pp. 581-586.
17. C. Van der Vleuten, Making the best of the 'long case', *The Lancet* **347** (1996), p. 704.
18. A. Dugdale, Long-case clinical examinations, *The Lancet* **347** (1996), p. 1335.
19. L. Olson, The ability of a long-case assessment in one discipline to predict students' perfomances on long-case assessments in other disciplines, *Academic Medicine* **74** (1999), pp. 835-839.
20. J. Price and J. Byrne, The direct clinical examination: an alternative method for the assessment of clinical psychiatry skills in undergraduate medical students, *Medical Education* **28** (1994), pp. 120-125.
21. V. Wass, R. Jones and C. Van der Vleuten, Standardized or real patients to test clinical competence? The long case revisited, *Medical Education* **35** (2001), pp. 321-325.
22. N. Teoh and F. Bowden, The case for resurrecting the long case, *British Medical Journal* **336** (2008), p. 1250.
23. J. Norcini, L. Blank, D. Duffy and G. Fortna, The mini-CEX: a method for assessing clinical skills, *Annals of Internal Medicine* **138** (2003), pp.476481http://www.biomedcentral.com/1472-6920/9/68

Education in Medicine Journal
2012, VOL 4 ISSUE 1
DOI:10.5959/eimj.v4i1.9

Appendix: Summary of long case evidence base

Author, date and country	Study group	Study type	Long case format	Key results	Study weakness
Chierakul *et al*, 2010, Thailand	585 postgraduate students sitting long cases for membership of the Royal College of Physicians of Thailand	Prospective cohort, multi-centre.	75 minute history and examination observed by 2 examiners. Examiners spend 20 minutes assessing patient first. Exam performed as a mid-year assessment and again as a final-year assessment.	Inter-rater reliability ranged from 15.3% to 27.3%. Significant correlation between mid-year and end-of-year scores.	Structure of examination changed during observed period (2005-2007) from 1 long case at final year assessment to 2 in 2007. Unclear if formal training received by examiners.
Wilkinson *et al*, 2010, Australia	59 postgraduates in two centres preparing for the entrance exam for the Royal Australasian College of Physicians clinical exam undertook 256 practice long cases.	Prospective cohort, two centre.	60 minutes of unobserved time with the patient for history and examination followed by 25 minutes presenting with 2 examiners. Each patient was assessed prior to the exam by the 2 examiners for 40 minutes. Each candidate performed 2 long cases with a different pair of examiners. Examiners received formal training.	Correlation (r=0.46) was demonstrated between the practice and examination long cases. The reliability of a single long case was greater under examination conditions than practice conditions.	Small numbers. No evaluation of validity.
Wilkinson *et al*, 2008, Australia	773 long cases undertaken as postgraduate examination for the Royal Australasian College of Physicians	Observational, multi-centre.	60 minutes of unobserved time with the patient for history and examination followed by 25 minutes presenting with 2 examiners. Each patient was assessed prior to the exam by the 2 examiners for 40 minutes. Each candidate performed 2 long cases with a different pair of examiners. Examiners received formal training.	38% of result variation due to candidate ability, 37% due to candidate x case interaction, 10% due to candidate x examiner interaction	Retrospective analysis. Little discussion of possible alternative examinations.
Hamdy *et al*, 2003, Bahrain	56 final year undergraduates.	Prospective cohort, single centre.	4 long cases consisting of 30 minutes observed history and examination and 15 minutes questioning by 2 examiners. 1 specialist, 1 non-specialist. Each panel of 2 examiners assessed 2 of the long cases for the same student. Each case selected from a specific blue-print. Student evaluated using a structured checklist. (DOCEE). Every three consecutive candidates examined with same set of patients and examiners.	Generalisability coefficient of 0.84 for 4 observed long cases with 2 examiners. Generalisability coefficient only theoretically increased from 0.61 to 0.62 when examiner number was increased from 10 to 50.	Single centre, small numbers.

Wass *et al*, 2001, UK	428 long cases undertaken by 214 final year medical school candidates	Observational, single centre.	14 minutes of observed history taking by 1 or 2 examiners followed by 7 minute presentation. Repeated twice with different examiners.	Little inter-case variation (long case 1 mean score 67.3%, long case 2 67.9%). 8-10 cases needed to achieve reliability >0.8.	Single centre. Inter-examiner variation not assessed. Unclear if formal training received by examiners.
Wass *et al*, 2001 UK	155 final year student long cases.	Observational, single centre.	16 minutes of observed history taking by 1 or 2 examiners followed by 8 minute presentation to 1 or 2 different examiners.	Inter-rater reliability correlations higher for observation (checklist 0.72 and global 0.71) than for presentation (checklist 0.38 and global 0.60)	Single centre. Unclear if formal training received by examiners.
Olson *et al*, 2000, Australia	67 undergraduate students in their fourth year undertaking one long case.	Randomised trial.	1 hour unobserved history and examination followed by 30 minutes for presentation and questions by 2 examiners. Students and assessors randomised to either to 'usual practice' or using a 'Structured Question Grid' whereby the examiners assessed the patient in advance and wrote down points they wished to examine.	No difference in chance of students being assessed as failing or of a discrepancy between students' and assessors' ratings of students as passing or failing. Students whose assessors used the grid felt it less representative of their ability.	Single centre.
Luiz *et al*, 2000, Brazil	27 undergraduates in the final two years of study.	Prospective cohort, single centre.	2 long cases consisting of 25 minutes observed history and examination with 8 minutes of structured questioning by 1 examiner. Observation of history and examination was scored using a 10 item checklist. Each case has a different examiner.	Agreement between different examiners on assessment of the various skills of the same student was 89%.	Single centre, small numbers.
McKinley *et al*, 2000, UK	19 postgraduates preparing for the entrance exam for the Royal Australasian College of Physicians. 9 had received 6 months exam training (group A), 10 had just started (group B).	Prospective cohort, single centre.	60 minutes of unobserved time with the patient for history and examination followed by 25 minutes presenting with 1 examiner.	Group A had a higher pass rate than group B.	Small numbers, single centre, no statistical analysis.
Olson, 1999, Australia	391 fourth or fifth year students sitting 1,564 long cases.	Prospective cohort, single centre.	40 minutes unobserved history taking and examination for surgery and reproductive medicine, 1 hour for internal medicine and paediatrics. 25 minutes to present to two examiners.	A passing grade on a single long case in any discipline reliably identified a majority of students as having a <10% risk of poor performance in any other discipline.	Single centre. Assessment not blinded. Unclear if formal training received by examiners.

				Little case variability.	
Abouna *et al*, 1999, Bahrain	74 final year undergraduates.	Prospective cohort, single centre.	4 to 6 long cases observed by 2 panels of 2 or 3 examiners – a mixture of specialists and non-specialists. Each case selected from a specific blue-print. Student evaluated using a structured checklist. Every three consecutive candidates examined with same set of patients and examiners. (IDOCEE).	Students and examiners were 'highly satisfied' with this form of examination.	Single centre. No assessment of inter-case or inter-rater reliability.
Price *et al*, 1994, Australia	178 undergraduate students.	Prospective cohort, single centre.	20 minutes history taking observed by 2 examiners followed by a case-specific task and 10 minutes of questioning.	High inter-rater reliability (kappa coefficient of 0.7).	Single centre, does not address generalisability. No inter-case analysis.
Newble, 1991, Australia	39 undergraduates in their final year.	Prospective cohort, single centre.	2 long cases consisting of 40 minutes of observed history and examination by 2 examiners. 10 minutes for presentation.	Students valued the assessment very highly (8.1/10).	Single centre, small numbers, no statistical analysis.

Corresponding author

Mr Simon Thornton
Department of Learning and Research,
Southmead Hospital, Bristol, UK BS10 5NB
Email: simon.thornton@nbt.nhs.uk

Accepted: Feb 2012 Published: June 2012