

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/8210250>

The long case

Article in *Medical Education* · December 2004

DOI: 10.1111/j.1365-2929.2004.01985.x · Source: PubMed

CITATIONS

44

READS

1,487

2 authors:



Val Wass

Keele University

133 PUBLICATIONS 4,044 CITATIONS

[SEE PROFILE](#)



Cees Van der Vleuten

Maastricht University

765 PUBLICATIONS 26,559 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Elaboration during problem-based, small group discussion [View project](#)



Computer learning [View project](#)

The long case

VAL WASS¹ & CEES VAN DER VLEUTEN²

BACKGROUND The long case has been gradually replaced by the objective structured clinical examination (OSCE) as a summative assessment of clinical skills. Its demise occurred against a paucity of psychometric research. This article reviews the current status of the long case, appraising its strengths and weaknesses as an assessment tool.

ISSUES There is a conflict between validity and reliability. The long case assesses an integrated clinical reaction between doctor and real patients and has high face validity. Intercase reliability is the prime problem. As most examinations traditionally used a single case only, problems of content specificity and standardisation were not addressed.

DISCUSSION Recent research suggests that testing across more cases does improve reliability. Better structuring of tests and direct observation increases validity. Substituting standardised cases for real patients may be of little benefit compared to increasing the sample of cases.

CONCLUSIONS Observed long cases can be useful for assessment depending on the sample size of cases and examiners. More research is needed into the exact nature of intercase and interexaminer variance and consequential validity. Feasibility remains a key problem. More exploration of combined assessments using real patients with OSCEs is suggested.

KEYWORDS education, medical undergraduate/
*methods; educational measurement/standards; clinical competence/standards; reproducibility of results.

¹School of Primary Care, University of Manchester, Manchester, UK

²Department of Educational Development and Research, University of Maastricht, Maastricht, The Netherlands

Correspondence: Val Wass, School of Primary Care, Rusholme Health Centre, Walmer Street, Manchester M14 5NP, UK.
Tel: 00 44 161 256 3015 (ext 231); Fax: 00 44 161 256 1070; E-mail: valerie.wass@man.ac.uk.

Medical Education 2004; **38**: 1176–1180

doi:10.1111/j.1365-2929.2004.01985.x

INTRODUCTION

The search for the ideal mode of assessment of clinical competence for undergraduates, which is both valid and reliable, remains controversial. Having been increasingly replaced by objective structured clinical examinations (OSCEs) throughout the world, the long case is still mourned,^{1,2} and, perhaps, rightly so. The tensions that exist between the validity and reliability of this assessment method and the feasibility of its delivery are difficult to resolve, but are similar to those experienced with any other form of assessment. Adequate sampling across a range of content is essential for any test of competence. Yet, the long case has educational advantages and, as more focus is placed on performance-based assessment, can be undertaken in the workplace. This article balances the strengths and weaknesses of the long case and argues for more research in this area.

VALIDITY OF THE LONG CASE

The American educationalist Flexner (1910) stated: 'There is only one sort of licensing test that is significant, i.e. a test that ascertains the practical ability of the students confronting a concrete case to collect all the relevant data and to suggest the positive procedures applicable to the conditions disclosed.'³ In the traditional long case, candidates are given uninterrupted and unobserved time, usually 30–45 minutes, to interview and examine a patient who has been selected from the wards or outpatient departments and who has had no training for examinations. Candidates then present their findings to the examiners as in an unstructured oral examination. The long case attempts to assess the integrated interaction between the doctor and a 'real' patient. An important aspect of the validity of a clinical examination is its

Overview

What is already known on this subject

Long cases are integrated in-depth assessments of clinical competence.

Observation adds to the reliability of the long case.

As for any measure of clinical competence, a single long case is unreliable.

What this study adds

Extending tests on long cases may successfully improve reliability.

Substituting standardised for real patients may add relatively little benefit.

Logistically, the resources required for long case testing limit its use.

Suggestions for further research

Future research should examine the relationships between intercase and interexaminer variance, the impact of using real rather than simulated patients, consequential validity, and the feasibility of combining long cases with other assessments.

approximation to the real world. The long case has arguable validity in this respect because the assessment is based on a highly authentic task and comes very close to a candidate's actual daily practice. Arguably, it is more valid than the tasks given to a candidate in a simulated and standardised situation such as in an OSCE. Studies investigating the construct validity are, however, lacking.⁴ Moreover, little is known about the consequential validity of the long case in terms of its impact on students' learning and preparation for examinations. Does the long case have greater consequential validity compared to the OSCE? This in itself is an interesting research question.

RELIABILITY OF THE LONG CASE

Why has the long case fallen from favour? Interrater reliability is not the major factor.⁵ The problem lies

with intercase reliability. Content specificity is now widely recognised as the most crucial issue in the assessment of clinical competence.^{6,7} Doctors and students do not perform consistently from task to task.⁸ A good performance on a single long case would not predict a good performance on another.^{9,10} Broad sampling across cases is essential to assess clinical competence reliably. Given the logistics of long case assessments, medical schools traditionally assessed students on a single case only. Implicit in this was the, perhaps rather naïve, assumption that experienced doctors had the skills to immediately identify good or weak students on a single patient interaction, and that this was predictive of any patient interaction. It is not surprising therefore, once the importance of context specificity was realised, that both undergraduate and postgraduate clinical assessments have moved to the multistation format of the OSCE.¹¹ However, concerns remain that by developing clinical assessments across a large number of contexts, little time is available on each case to fully assess the candidate's competence. The validity of the OSCE is being questioned in this respect.¹² Depth of assessment, as argued for by Flexner, has been lost in order to gain breadth. Thirty years after its introduction, life beyond the OSCE is now under discussion.¹³

We are looking for ways forward. Surprisingly, the move away from the long case occurred in the face of very little published psychometric evidence.⁴ Yet the long case does have face validity and it would be unwise to abandon this form of testing without clearer statistical evidence of how it performs as a test. Can the reliability and validity of the long case be improved?

IMPROVING THE INTERCASE RELIABILITY OF THE LONG CASE

Increasing the number of long cases seen by each student should, theoretically, address the issue of content specificity. If infinite resources were available, how many cases would be necessary to achieve the intercase reliability needed for a high stakes assessment? By comparing final year medical student performance across 2 observed, modified history taking long cases using 2 pairs of different examiners we predicted, using generalisability theory, that 10 cases would achieve reliability of 0.8 (Cronbach's alpha).¹⁴ The calculation assumed different examiners were used for each case. Thus, a large sample of examiner judgements was also achieved. In terms of

the testing time required for this increased sample, the reliability outcome is no better or worse than for any other measure of clinical competence.⁶ More studies are needed to both replicate this finding and investigate the relative magnitude of case and examiner variance. So far there is no reason to believe that, provided sufficient cases and examiners are used, the long case would differ significantly from, for example, the OSCE.^{15–17} Given sufficient testing time and a large patient and examiner resource, a reliable high stakes long case examination theoretically has potential.

The key difference between the long case and the OSCE is the unstandardised nature of the patients. Long case examinations can never be equivalent across a cohort of candidates. But does this matter? Efforts to standardise encounters and not use real patients may lead to relatively small gains compared to ensuring that sufficient encounters are assessed to overcome the problems of content specificity.^{18,19} Logistically, this remains difficult. Hamdy *et al.* from Bahrain recently demonstrated that a 3-hour examination of 4 45-minute observed long cases had good reliability.²⁰ Real patients selected from a predetermined blueprint of common disease were used. In the USA, the mini-CEX (mini-clinical evaluation exercise) work-based assessments use limited observation of the history and examination of real patients to assess clinical competencies.²¹ Durning *et al.* reported acceptable reliability across 7 such real patient cases.¹⁶ These findings continue to challenge the assumption that standardisation of cases is essential for the reliable assessment of clinical competence.

Standardising patients does have great advantages. Real patients can be a liability.²² Standardisation enables accurate blueprinting of the test. Yet it requires a high level of training and resource. This is feasible in some postgraduate examinations, where the cost can be covered by candidate fees, but it remains difficult in many undergraduate universities and countries with limited resources. Simulation moves the assessment away from the workplace. In our increasingly diverse society, it is difficult to create simulations that mirror the range of ethnicity. Circumstances such as those involving patients with limited language, the need to use interpreters, limited cultural understanding between doctor and patient, etc., present complex challenges for standardisation, which might be best addressed using a variety of real cases. As we strive for maximum authenticity, research to improve our understanding in this area is needed. Standardising encounters may not impact on reliability as much as was originally assumed. Content

specificity appears to be the key issue and ‘noise’ associated with the authenticity of the patient presentation seems to subordinate this effect.

IMPROVING THE VALIDITY OF THE LONG CASE

Over the years attempts have been made to improve the validity of the long case format by increasing the authenticity of the assessment. It would seem logical that, rather than relying on a presentation alone, observation of the candidate while eliciting the history and carrying out the examination would be a more valid assessment of the candidate’s competencies. The use of observed long cases has been reported in some institutions.^{23,24} Gleeson developed a more structured presentation of an unobserved long case, the objective structured long examination record (OSLER),²⁵ which includes some direct observation of the candidate interacting with the patient. Fraser developed it as a formative tool for assessment of both undergraduates and postgraduates within the Leicester Assessment Package.²⁶

A key question concerns whether observation adds to the validity of the assessment.

A recent study demonstrated that observation does measure a useful and distinctive component of history taking clinical competence over and above the contribution made by the presentation. We observed an undergraduate history taking long case and compared results of the observed and presentation component with performance on an OSCE undertaken at the same time.⁵ More studies are needed to investigate the construct validity of such improvements of the long case examination.

FUTURE RESEARCH

This review raises more questions than it answers. It describes the current state of play of the long case. More research into the psychometrics of the long case is required. There is a wide range of literature available on the reliability of the OSCE. We need more information on the intercase and interexaminer variance of the long case. A key question concerns how different structures for the long case affect these variances. Whether the long case survives or vanishes from the assessment scene should not merely be based on opinionated arguments but on evidence originating from appropriate research.

Almost 100 years after Flexner's observation, the long case continues to have undoubted validity. Critics of the OSCE may be justified in their concerns about the failure of this examination format to integrate the whole process of clinical assessment, from history taking through to the management of a particular case. The logistics of providing a sufficient number of long cases to achieve a reliable high stakes summative test of clinical competence challenges its feasibility. We encourage its use in low stakes (formative) assessment with emphasis on the importance of including observation in the process. Alternatively, a compromise could be reached if 1 or more long cases were to be combined with OSCE stations, to bring both depth and breadth to the assessment. But again, any such moves should be underpinned by good research.

It is time to fill the gaps in the long case literature. We need to know more about:

- *Intercase variance*: how many observed full long case assessments are necessary to produce a reliable assessment of clinical competence?
- *Interexaminer variance*: how much do examiners contribute to the reliability and, provided sufficient cases are tested, how many different examiners are needed?
- *Real patient variance*: to what extent does the use of real patients affect the reliability?
- *Construct validity*: what is the effect of making the long case as authentic as possible? What is the incremental validity of using real patients versus standardised patients? How can the long case examination be combined with other formats?
- *Consequential validity*: what sort of educational consequences does a (reliable) long case have when compared to the OSCE?

In the meantime, provided the stakes are not high, long cases remain a useful tool for teachers to observe students in action and give feedback. For summative purposes, issues of case specificity and effective blueprinting within the logistics of running a series of long cases confound feasibility. We would not recommend the use of the long case in high stakes assessments of clinical competence for this reason.

CONTRIBUTORS

Both authors reviewed and appraised the current literature, discussed areas for further research and contributed to the writing of this article.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICTS OF INTEREST

None.

ETHICAL APPROVAL

Not applicable.

REFERENCES

- 1 Meadow R. The structured exam has taken over. *BMJ* 1998;**317**:1329.
- 2 Fraser R. Does observation add to the validity of the long case? [Letter.] *Med Educ* 2001;**35**:1131–3.
- 3 Flexner A. *Medical Education in the United States and Canada*. Bethesda, Maryland: Science and Health Publications 1910.
- 4 van der Vleuten CPM. Making the best of the 'long case'. *Lancet* 1996;**347**:704–5.
- 5 Wass V, Jolly B. Does observation add to the validity of the long case? *Med Educ* 2001;**35**:729–34.
- 6 Wass V, van der Vleuten CPM, Shatzer J, Jones R. Assessment of clinical competence. *Lancet* 2001;**357**:945–9.
- 7 van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ* 1996;**1**:41–67.
- 8 Swanson DB, Norman GR, Linn RL. Performance-based assessment: lessons learnt from the health professions. *Educ Res* 1995;**24** (5):5–11.
- 9 Norcini JJ. The death of the long case? *BMJ* 2002;**324**:408–9.
- 10 Norcini JJ. Does observation add to the validity of the long case? [Letter.] *Med Educ* 2001;**35**:1131–3.
- 11 Harden RM, Gleeson FA. ASME Medical Educational Booklet, 8. Assessment of medical competence using an objective structured clinical examination (OSCE). *J Med Educ* 1979;**13**:41–54.
- 12 Hodges B. Validity and the OSCE. *Med Teacher* 2003;**25**:250–4.
- 13 Ben-David MF. Life beyond the OSCE. *Med Teacher* 2003;**25**:239–40.
- 14 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.

- 15 Petrusa ER. Clinical performance assessments. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook for Research in Medical Education*. Dordrecht: Kluwer Academic Publisher 2002;673–709.
- 16 Durning SJ, Cation LJ, Markert RJ, Pangaro LN. Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. *Acad Med* 2002;**77**:900–4.
- 17 Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med* 2003;**15**:270–92.
- 18 Norman G. The long case versus objective structured clinical examinations. *BMJ* 2002;**324**:748–9.
- 19 Norcini JJ. The validity of long cases. *Med Educ* 2001;**35**:135–7.
- 20 Hamdy H, Prasad K, Williams R, Salim FA. Reliability and validity of the direct observation clinical encounter examination (DOCEE). *Med Educ* 2003;**37**:205–12.
- 21 Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Intern Med* 1995;**123**:795–9.
- 22 Sayer M, Bowman D, Evans D, Wessier A, Wood D. Use of patients in professional medical examinations: current UK practice and the ethicolegal implications for medical education. *BMJ* 2002;**324**:404–7.
- 23 Newble DI. The observed long case in clinical assessment. *Med Educ* 1994;**25**:369–73.
- 24 Price J, Byrne GJA. The direct clinical examination: an alternative method for the assessment of clinical psychiatric skills in undergraduate medical students. *Med Educ* 1994;**28**:120–5.
- 25 Gleeson F. The effect of immediate feedback on clinical skills using the OSLER. In: Rothman AI, Cohen R, eds. *Proceedings of the Sixth Ottawa Conference of Medical Education*. Toronto: University of Toronto Bookstore Custom Publishing 1994;412–5.
- 26 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.

Received 2 February 2004; editorial comments to authors 10 March 2004; accepted for publication 21 June 2004