Education in Medicine Journal

SPECIAL COMMUNICATION

Title: Long Case Examination As Practiced in Malaysian Medical Schools: Are We Doing It Right?

Authors: Norwati Daud, Nurulhuda Mat Hassan, Nurul Izza Yunus

Submitted Date: 25-10-2022

Accepted Date: 14-02-2023

Please cite this article as: Daud N, Mat Hassan N, Yunus NI. Long case examination as practiced in Malaysian medical schools: are we doing it right? Education in Medicine Journal (Early view)

SPECIAL COMMUNICATION

ARTICLE INFO

Submitted: 25-10-2022
Accepted: 14-02-2023

Long Case Examination As Practiced in Malaysian Medical Schools: Are We Doing It Right?

Norwati Daud, Nurulhuda Mat Hassan, Nurul Izza Yunus

Faculty of Medicine, Universiti Sultan Zainal Abidin, Kuala Terengganu, Terengganu, Malaysia

To cite this article: Daud N, Mat Hassan N, Yunus NI. Long case examination as practiced in Malaysian medical schools: are we doing it right? Education in Medicine Journal (Early view)

ABSTRACT

'Long case' examination has long been practised as part of the clinical component assessment in many medical schools in Malaysia. The most important aspect of an evaluation is its validity, which should meet the criteria for a good evaluation. Under-sampling, unobserved, unstructured rating and case specificity are all issues with the type of long case examination that is currently used in Malaysia, and they can all have an impact on its overall validity. The validity of the current conduct of the long case examination can be argued especially from the aspect of testing the 'Show How' and 'Does' component. Long case examination can be improved by increasing the number of cases, observing the examination and providing a more objective rating. However, this solution may lead to some logistic issues. In conclusion, the approach to planning for a long case examination should consider the utility of the examination which is the sum of validity, acceptability, reliability, feasibility, and educational impact.

Keywords: *Long case, long consultation, medical students, assessment, validity*

CORRESPONDING

Nurul Izza Yunus, Faculty of Medicine, Universiti Sultan Zainal Abidin, Kuala Terengganu, Terengganu, Malaysia

Email: izzayunus@unisza.edu.my

INTRODUCTION

'Long case' examination is a common practice in assessing clinical competency in the final professional examination/exit examination in Malaysian medical schools. It is usually used with Observed Structured Clinical Examination (OSCE) and short cases. Through personal experience and communication, these three types of examination are usually compensatory for the clinical component. A description of a long case examination in the final professional examination in Malaysia is typical as below:

A student will be provided with one long-case patient, which is usually a real patient preselected from hospital-based or outpatient based. The student will either get a long case from medical-based (adult medicine, paediatric or psychiatry) or surgical-based (surgery, obstetrics, and gynaecology or orthopaedic). A student who gets a medical-based patient as a long case will then get surgical-based short cases (between 2 to 4 cases). During a long case, a student will spend one hour with a patient. In most cases in Malaysia, it is unobserved, where he/she is expected to get the medical history and

physical examination. After one hour, students will go to a separate room (without the patient), and they are expected to present the medical history and physical examination to a panel of examiners (between 2 to 3). There will be questions and answers by the examiners, usually in provisional and differential diagnoses, investigations, and management. On some occasions, examiners will take students back to the patient to confirm history or physical findings. Examiners are usually provided with the patient's summary or record if feasible. Besides, examiners sometimes have the opportunity to see the patient first to confirm some history and physical examination (from our experience, this is rarely done for long cases but is usually done for short cases). Otherwise, examiners will just rely on the patient's summary prepared by the examination committee or the patient's medical record.

Another typical feature of a long case examination in Malaysia is that the panel consists of different specialities i.e. for a medically based panel, it will consist of a physician, a paediatrician and a psychiatrist (or any of the two). For a surgical-based panel, it will consist of a surgeon, an obstetric and gynaecologist and an orthopaedic surgeon (or any of the two). For a medical-based case, if it is a paediatric case, the examination is usually led by the paediatrician, and the other two examiners play a minimal role. The same goes for surgical-based cases. Although marks are usually given as a consensus, it mostly relies on the judgment of the leader examiner. For example, if the case is surgical, the lead examiner would be the surgeon, and the other examiners from a different speciality (orthopaedic and/or obstetrics & gynaecology) would usually agree with the marks given by the surgeon.

CRITERIA FOR A GOOD ASSESSMENT

According to Ottawa 2010 conference consensus statement (Norcini et al., 2011), criteria for a good assessment include validity/coherence, reproducibility/consistency, equivalence, feasibility, educational effect, catalytic effect, and acceptability (1). Zubair & Khoo (2009) in his book stated that validity is defined as the extent to which an assessment instrument measures what it intends to measure (2). Reliability means the consistency of a test score. Meanwhile, feasibility refers to the overall ease of construction, administration, scoring, and reporting of the assessment. It may also include the cost of using the instrument. Educational impact means the consequences of an assessment on student learning and professional development, whether intentional or unintentional.

McAleer (2009) proposed five major types of validity namely, content validity, concurrent validity, predictive validity, construct validity, and face validity (3). Content validity is an assessment that is able to represent the content of a course. Construct validity refers to the extent to which the test measures an intended construct. The assessment method should be suitable to the nature of what it is intended to assess. Predictive validity means how predictive the test is to the performance in other areas/situations or future performance. Face validity means that the test should appear as though it measures what it is supposed to measure. Concurrent validity is the degree to which the test score correlates with other established test scores measuring the same construct.

Downing and Haladyna (2009) stated that there are five major sources of test validity evidence (4). They are content, response process, internal structure i.e, reliability, relations to other variables, and evidence based on consequences of testing.

THE ISSUE

Is our 'so-called long case' really testing what we want to test?

This is probably the most important initial question. "What do we want to measure/test from a long case?". This is also a question that we must ask when we want to select an assessment tool.

It is assumed that the long case intends to assess the 'show how' and 'does' levels in Miller's pyramid. Therefore, it is a combination of observation and performance assessment. Observation assessment is defined as an 'interpersonal activity and is subject to all of the potential pitfalls of human relation' (5). A performance assessment is an examination designed to elicit performance on an actual or simulated real-life task (6). While performing performance testing, the entire process should be observed. Here comes the issue of how long the case is carried out in Malaysia, that it is unobserved. The checklist and rating scale provided to the examiner usually consists of communication skills and examination techniques. This is contrary to how the long case is carried out. The medical history and physical assessment findings rating is based on what students present and not how students elicit in the history and perform in physical examination. So, the rating is just based on presentation than skills of doing (does). These skills would be communication and clinical examination skills/techniques. The examiner's rating assumes that if a student can present relevant and correct history, he/she should have good communication skills. If a student can elicit the correct physical findings, he or she has used the proper techniques. This is not always the case. Many doctors overestimate their skills, and self-assessment is usually inaccurate (7). Therefore, skills, especially communication, should be observed or at least recorded and reviewed (8).

The practice of unobserved long cases may affect its validity. For example:

1. Language bias
   Since English is the main communication medium in medicine, students whose English language is good may look impressive with the language, which may affect the rating.
2. Halo effect
   Students with some interesting character or attitude may affect their rating.
3. Students may cheat i.e., they may present history and physical findings the way they assume correct and not the true history or findings. The content in the history might not be the same in the presentation, and students can even make up some information.

The second part of the long case examination is the discussion of the case i.e., investigation and management. This part of the assessment is valid without observation. However, this part is mainly assessing students' cognitive and problem-solving skills.

Is a long case a valid assessment tool?

Wass and van der Vleuten (2004) agreed that a long case has high authenticity and is probably more valid than an OSCE (9). The advantage of a long case is it reflects a student's

performance with a real patient. Although the literature has shown that OSCE has higher validity, OSCE is somewhat artificial, where patients are usually simulated patients, and the examination setting usually takes place outside the ward. Communicating with a real patient is more difficult than communicating with a simulated patient, where the script is usually prepared and the patient's attitude is an act. Similarly, an actor patient cannot always simulate an examination with real patients. As a result, the long case may have a high level of content and construct validity. Despite the ability of the long case to represent the real case, the low validity of the long case is because of its number/sampling. Most students will get one long case (rarely two in a Malaysian setting). Long case tried to generalise the candidate's performance across all clinical problems based on one long case (11, 22, 10, 9). Therefore, it is difficult to conclude that performance in just one case reflects the true performance or consistency in performance if the student is given another case. Hence, the strength of the OSCE is the sampling of many competencies, whereas its downside is compartmentalization. The strength of the long case is authenticity, whereas its downside is small or no sampling.

As discussed above, our long case is mainly unobserved. Wass & Jolly (2001) found that inter-rater reliability among the examiners for observed long case history taking is much higher than for presentation only (0.72 vs 0.38) (10). The correlation between the observed history-taking and presentation is low (0.38).

Is a long case a reliable assessment tool?

The reliability of the long case has long been argued. Wilson et al., (1969) described that marks by a single examiner have a poor correlation with marks given by two examiners on the same student (12). Marks by the individual examiner have poor consistency when given to the same student at different times. Moreover, Norcini (2002) stated that if a single long case is used, the reliability is low, 0.24 (1). Low reliability is affected by case specificity, examiners' differences in marking and differences in areas of competency assessed (1). The reason why reliability is low is that long case is case specific, examiners vary in expectation when the rating is unstructured, and cases are different in difficulties. The single case leads to under-sampling and hence is a threat to the validity of the assessment. A study by Wilkinson et al., (2008) found that to achieve a reliability of 0.8, it requires five to six 85-minute-long cases (13). Their study also showed that, with one long case without any short case, reliability is 0.43. However, if one long case is combined with 3 short cases, reliability is increased to 0.59 (similar to what is practised in Malaysia). According to the article, to achieve at least 0.8 for a high-stake examination, a combination of 3 long cases and 7 short cases will achieve the value. Norman (2002) stated that if 10 observed long cases with a structured rating and 30 OSCEs are used, the reliability is equivalent (14). However, it is almost impractical to use 10 long cases. Another study by Hamdy et al., (2003) showed that the generalisability coefficient with four cases and two raters was 0.84 (15). Increasing the cases from one to four improved reliability to above 0.8. However, increasing the number of raters had little impact on reliability.

The next aspect being considered is the correlation between long cases and OSCE. If the long case and OSCE measure the same construct, the correlation between the long case and OSCE should be high. An analysis of a professional examination at Universiti Kebangsaan Malaysia showed that the correlation between long cases and OSCE is very weak, 0.17 (16). A similar

analysis of the Universiti Sultan Zainal Abidin professional examination showed a fair correlation between the OSCE score and long case score, 0.48 (17).

Another element that should be considered is predictability. A study by Olson (1999) showed that performance in the long case of one discipline is a good predictor for performance in other disciplines (18). Another study by Probert (2003) found that OSCE was more consistently associated with positive ratings by the consultant among house officers (19). The long case was inversely associated with the consultants' report.

Observation is important in the long case as students tend not to focus on physical examination skills and focus more on history (23). Besides, the observed long case provides higher reliability (0.70 vs 0.38) (10). However, since the observed long case is not a common practice in Malaysia, how would students feel if we introduce an observed long case? Newble (1991) investigated students' perceptions of observed consultation during the initial planning for the change (20). Students thought it was extremely beneficial but not enjoyable. Although non-participating students in this study had similar perceptions, they were less enthusiastic about the idea.

HOW TO IMPROVE A LONG CASE EXAMINATION?

Michael et al., (2013) recommended that the long case should use the strength of an OSCE, which are (28):

1. Structuring the format and the marking scheme so that the competencies assessed are uniform and the rating is more objective.
2. Increase the number of examiners.
3. Observing the long case.
4. Increase the number of cases.
5. Make the long case shorter (20-45 minutes).

Gleeson (1997) recommended OSLER (Objective Structured Long Examination Record), where history and physical examination are observed, and the examiner's checklist and rating are structured (29). Structured examiner marking is useful for making questions more uniform. However, it does not affect the chance of passing or failing (24). Besides, the difference in rating based on case-specific tasks e.g., difficulty level may increase the inter-rater reliability (25).

If ten cases are used, each of 20-minute duration, it will give rise to a reliability of between 0.84 to 0.88 (27). Among the conclusion by Ponnamperuma et al., (2009) from a literature review on a long case, two long cases with two examiners are effective from student and staff perspectives (20). Examiner agreement could achieve 89% if students took 2 long cases, which are structured and observed, and with different examiners (26). Norcini (2002) suggested adding more short cases and OSCE stations to increase the validity of the clinical component (1). He added that increasing the number of examiners will have a great effect, although beyond 4 to 5 examiners will not give further improvement.

If we were to apply the recommendations to improve the validity of a long case, what will be the issue/problem?

Van der Vleuten (1996) suggested an approach for an assessment in which he used utility as a mathematical sum (11):

Utility = validity x acceptability x reliability x feasibility x educational impact

The most critical issue/problem is most likely feasibility. As previously stated, feasibility influences utility since the observed long case necessitates more time for the examiners. Increasing the number of examiners necessitates an increase in human resources as well as a budget increase. Increasing the number of cases requires more time and more cases, which necessitates a larger budget and more space.

Although this article is not meant to discuss cost, to be specific, the cost is an important issue to decide on feasibility. Take a few medical schools in Malaysia as examples, with medical students' enrolment between 60 to 250 students per batch. (30)(31)(32). Preparing for, e.g., the final examination is costly, especially in providing an honorarium to patients, examiners, and extra staff. Currently, most universities are paying about RM 100.00 per patient for exams. Assuming that for a long case examination, 1 patient is used for 4 candidates, then for 200 candidates, about 50 patients are needed and consequently, the cost would be about RM 5,000.00. The cost is doubled if two long cases are used. This does not include patient meal costs. Not only is it expensive, but finding suitable patients for long cases is also difficult. Handling patients for examination is a difficult task. A large number of people are required for the process. Long case examinations are typically conducted alongside short case examinations, with at least three patients for every four or five students. So it costs roughly the same as a single long case. With a single long case, the maximum number of candidates that can be handled daily is about 30 to 50 students, depending on how many examiners are available. For 200 students, the long case (together with the short case) will last about 1 week of working days. Two long cases may mean 2 weeks of work. With limited staff and examiners, it will be very exhaustive.

From the literature, a minimum of two examiners are needed for each student to have acceptable reliability but beyond four has no added value. A minimum of three long cases, if combined with seven short cases, will achieve favourable reliability. Considering the aforementioned issues, it is probably not that feasible.

Another important issue in the practice of long case examination in Malaysian medical schools is the alternate medical-surgical based on long-short case arrangement. A student who gets a medical-based long case will get a surgical-based short case. Assume a student is assigned a psychiatry case as his or her long case (medical-based). The student will then be assigned a surgical-based short case that includes surgery, obstetrics, gynaecology, and orthopaedics. The student is not assessed in any of the medical competencies (internal medicine and paediatric), which are the bread and butter of medical principles and practice. If a student gets an orthopaedic case for a long case, the student is not assessed at all in any general surgery competency. Roughly about one-third of students may fall into this category of not being assessed at all in the area of medicine and general surgery. It should be noted

that general surgery and medicine are very important specialities during houseman ship training. How do we ensure the student is competent in these two important specialities?

CONCLUSION

Studies showed that a long case has good content validity as it reflects the real case, and its authenticity is undisputed. However, the issue with a long case is with its under-sampling, unobserved, unstructured rating, and case-specific, which all can affect its validity as a whole. Recommendations to improve the validity of long cases by increasing the number of cases, making it an observed long case, and providing structured rating so that the competency being assessed is clear uniform, and objective. The long case should not be used to evaluate clinical competence on its own. It should be used in conjunction with other methods, such as short cases and OSCE. However, carrying out the recommendations may be prohibitively expensive and time-consuming. A cost-benefit or, more precisely, cost-validity judgement is critical. Validity is important because it is part of a high-stakes examination, but it should also be feasible to achieve the most acceptable standard.

REFERENCES

1. Norcini, J.J., (2002). The death of the long case? BMJ 324, 408–409. doi:10.1136/bmj.324.7334.408

2. Zubair, A., Khoo, H. E., (2009). 'How Do We Assess?' in Basics in Medical Education. 2nd edn. World Scientific. Chapter 24.

3. McAleer S., (2009) 'Choosing assessment instrument' in Dent J.,A., Harden R., M. (ed). A Practical Guide for Medical Teachers. 3rd edn. Churchill Livingston Elsevier, p.318.

4. Downing, S.M., Haladyna T.M., (2009) 'Validity and its Threats' in Downing M.D., Yudkowsky R. (ed) Assessment in Health Professions Education. Routledge, p. 29.

5. McGaghie W.C., Butter J., Kaye M. (2009) "Performance tests' in Downing M.D., Yudkowsky R. (ed) Assessment in Health Professions Education. Routledge, p. 192.

6. Yudkowsky, R., (2009) "Performance tests' in Downing M.D., Yudkowsky R. (ed) Assessment in Health Professions Education. Routledge, p. 232.

7. Davis, D.A., Mazmanian, P.E., Fordis, M., et al., (2006). Accuracy of physician self-assessment compared with observed measures of competence: A systematic review. Journal of the American Medical Association. doi:10.1001/jama.296.9.1094

8. Ha, J.F., Longnecker, N., (2010). Doctor-patient communication: a review. The Ochsner journal 10, 38–43. doi:10.1043/toj-09-0040.1

9. Wass V., van der Vleuten C. (2004), The long case. Medical Education, 38, 1176–1180.

10. Wass, V., Jolly, B., (2001). Does observation add to the validity of the long case? Medical Education 35, 729–734. doi:10.1046/j.1365-2923.2001.01012.x

11. van der Vleuten, (1996). Making the best of the 'long case'. Lancet, 347 (3),704–705.

12. Wilson, G., Lever, R., Harden, R.M., et al., (1969). Examination of clinical examiners. The Lancet 293, 37–40. doi:10.1016/S0140-6736(69)90998-2

13. Wilkinson, T.J., Campbell, P.J., Judd, S.J., (2008). Reliability of the long case. Medical Education 42, 887–893. doi:10.1111/j.1365-2923.2008.03129.x

14. Norman, G., (2002). The long case versus objective structured clinical examinations. BMJ 324, 748–749. doi:10.1136/bmj.324.7340.748

15. Hamdy, H., Prasad, K., William, R., Salih, F.A., (2003). Reliability and validity of the direct observation clinical encounter examination (DOCEE). Medical Education 37, 205–212. doi:10.1046/j.1365-2923.2003.01438.x

16. Kamarudin, M.A., Mohamad, N., Awang Besar, M.N., et al., (2012). The relationship between modified long case and objective structured clinical examination (OSCE) in final professional examination 2011 held in UKM Medical Centre, Social and Behavioral Sciences 60, 241-248.

17. Husbani, M.A.R., Norwati, D. Moe, M., et al., (2017). Does OSCE performance correlate with long case performance in final professional MBBS examination, Universiti Sultan Zainal Abidin? 1st International Community Health Conference 2017, UniSZA, Terengganu.

18. Olson, L.G., (1999). The ability of a long-case assessment in one discipline to predict students' performances on long-case assessments in other disciplines. Academic medicine : journal of the Association of American Medical Colleges 74, 835–9.

19. Probert, C.S., Cahill, D.J., McCann, G.L., Ben-Shlomo, Y., (2003). Traditional finals and OSCEs in predicting consultant and self-reported clinical skills of PRHO: a pilot study. Medical Education 37, 597-602.

20. Newble, D.I., (1991). The observed long-case in clinical assessment. Medical Education 25, 369–373. doi:10.1111/j.1365-2923.1991.tb00083.x

21. Ponnamperuma, G.G., Karunathilake, McAleer, S., Davis, M.H., (2009). The long case and its modifications: A literature review. Medical Education. doi:10.1111/j.1365-2923.2009.03448.x

22. Dugdale A., (1996). Letters: long case clinical examinations. Lancet, 347, 1335.

23. Pavlakis N, Laurent R., (2001). Role of the observed long case in postgraduate medical training. International Medical Journal, 31(9), 523–528.

24. Olson LG, Coughlan J, Rolfe I, Hensley MJ., (2000). The effect of a structured question grid on the validity and perceived fairness of a medical long case assessment. Medical Education, 34(1):46–52.

25. Price J, Byrne JA., (1994). The direct clinical examination: an alternative method for the assessment of clinical psychiatry skills in undergraduate medical students. Medical Education, 28 (2), 120–125.

26. Luiz, E. A. T., Roberto, O. D., Fernando, C. F., Eduardo F, Lio, C.M., Ana, L. C. M., Lio, C. V., (2000). A standardised, structured long case examination of clinical competence of senior medical students. Medical Teacher, 22 (4), 380–385.

27. Wass, V., Jones, R., Van Vleuten, C.D., (2001). Standardized or real patients to test clinical competence? The long case revisited. Medical Education 35, 321–325. doi:10.1046/j.1365-2923.2001.00928.x

28. Michael, A., Rao, R., Goel, V., (2013). The long case: a case for revival? The Psychiatrist 37, 377–381. doi:10.1192/pb.bp.113.043588

29. Gleeson, F., (1997). AMEE Medical Education Guide No. 9. Assessment of clinical competence using the Objective Structured Long Examination Record (OSLER). Medical Teacher 19, 7–14. doi:10.3109/01421599709019339

30. Universiti Malaya, Convocation Ceremony 61, pp 44, [Internet], 2022 [cited 29th Jan 2023] https://umconvo.um.edu.my/convocation-e-book.

31. Universiti Teknologi Mara, Prospectus Faculty of Medicine, pp 5, [Internet], 2022 [cited 29th Jan 2023] https://medicine.uitm.edu.my/images/OFFICE/corporat-comm/PROSPECTUS-FACULTY-OF-MEDICINE-1st-EDITION-(14-SEPT-2021).pdf

32. Universiti Sains Malaysia, Bakal bergelar doctor perubatan, 113 pelajar PPSP dirai, [Internet] 2021, [cited 29th Jan 2023] http://www.kk.usm.my/index.php/news-media/1391-bakal-bergelar-doktor-perubatan-113-pelajar-ppsp-dirai