Quest International Journal of Medical and Health Sciences – [Online]

ORIGINAL ARTICLE

The "failure to fail" phenomenon in the clinical long case examination

Ganesh Ramachandran[1]*, Aung Ko Ko Min[2], Sarmishtha Ghosh[3]

Corresponding author:

[1]Assoc.Prof. (Dr.) Ganesh Ramachandran, Deputy Dean Academic Affairs, Faculty of Medicine and Biomedical Sciences, MAHSA University, Selangor, Malaysia

Email: ganesh@mahsa.edu.my

[2]Dr. Aung Ko Ko Min, Department of Community Medicine,Faculty of Medicine and Biomedical Sciences MAHSA University, Selangor, Malaysia

[3]Assoc Prof (Dr) Sarmishtha Ghosh, IMU School of Education, International Medical University Kuala Lumpur

Information about the article:

Received: July. 20, 2018
Accepted: June. 20, 2019
Published online: July 1, 2019

Cite this article:

Ramachandran G, Ko KMA, Ghosh S. The "failure to fail" phenomenon in the clinical long case examination. Quest International Journal of Medical and Health Sciences. [internet], 2019 [2019/7/1]; 2(1):3-7. Available from: http://www.qiup.edu.my/wp-content/uploads/2019-2.pdf

Publisher

Quest International University Perak (QIUP), No.227, Plaza Teh Teng Seng (Level 2), Jalan Raja Permaisuri Bainun, 30250 Ipoh, Perak Darul Ridzuan, Malaysia

e-ISSN: 2636-9478

ABSTRACT

Introduction:
In the final clinical examination for undergraduate students, real patients are used to assessing the clinical competence of the students. Numerical scores based on a predetermined rubric are awarded to decide the status of the students. Subjective comments by examiners are encouraged to assess concordance. The aim of the study is to find out if the subjective comments match the numerical scores for the students undertaking the examination.

Methods:
This is a cross-sectional study for a batch of 106 students. The mark sheet was framed with standard criteria for a long case and examiners were briefed to give marks according to the criteria; they were allowed to give free comments. These were collected and thematic analysis was conducted. The categories used were "do not tally", "tally" and "no comment".

Results:
In medicine and surgery, 0% – 4.5% of the responses did not tally, whereas 38%-50% tallied. In the "no comment" category, all except two candidates passed. In paediatrics and psychiatry, 35%-50% of the responses did not tally, while 25%-50% were with no comments but clear passes. Obstetrics and gynaecology (O&G) and orthopaedics had 18%-30% responses that did not tally, and 18%-40% responses that showed concordance. The significant observation was 0% of "do not tally" in surgery and 0% "tally" in psychiatry.

Conclusion:
The disparity of subjective comment was lowest in medicine and surgery whereas it was highest in psychiatry. Paediatrics, O&G and orthopaedics have a considerable concordance. Possible causes and solutions are discussed.

Keywords
Clinical assessment, concordance, discordance, long case subjective comments

Introduction

The clinical examination in the MBBS Final Professional Examination at the Faculty of Medicine, MAHSA University comprises clinical cases and an Objective Structured Clinical Examination made up of static and interactive stations. The clinical cases are picked by a panel of examiners from cases available in the wards and outpatient clinics. The disciplines tested are internal medicine, general surgery, obstetrics and gynaecology, orthopaedics, paediatrics and psychiatry. All students are examined by a pair of examiners using a pre-determined rubric that has been approved by the faculty academic board. The final marks for each student are consensual and examiners are required to provide subjective comments on the performance of each student with the marks allotted. A robust examination process in this portion of the clinical examination is essential, as this is an exit examination that allows the student to commence two years of house officer training prior to full registration with the Malaysian Medical Council. Apart from a numerical score, the examiners were required to provide a subjective comment on the performance of each student. The purpose of this study is to determine the level of concordance between the marks allotted and subjective comments in the clinical long case. It is felt that subjective comments would be a reflection of the global ability of the student. If this were so, the numerical score should mirror these subjective comments. There has been concern that borderline numerical scores are usually an indication of poor performance in the future [1] and may indicate reluctance on the part of examiner to fail students. As such, an additional rating mechanism would be useful in improving the reliability and validity of the assessment process, particularly in pass/fail decisions [1]. The study aimed to determine the association between the subjective comments and the numerical marks obtained using a predetermined rubric.

Methods

Study Period

A cross-sectional study was conducted in 2016 on the Final Professional Examination record of the graduating class of medical students in the 2014/2015 academic session.

Study design, participants and the collection of data

This examination had two parts: a theory portion and a clinical portion. The theory portion comprised two multiple choice question papers, two modified essay question papers, and two short essay question papers. The clinical portion comprised a clinical long case, three clinical short cases, and a fifteen station Objective Structured Clinical Examination. All students were subjected to one clinical long case from internal medicine, general surgery, obstetrics and gynaecology, orthopaedics, paediatrics or psychiatry. Internal medicine, paediatrics, and psychiatry were grouped as medicine and allied subjects. General surgery, orthopaedics, and obstetrics and gynaecology were grouped as surgery and allied subjects. Patients were selected for the examination using a pool of non-acute admitted patients and outpatients on follow up at the specialist clinics of the respective departments. Selection of patients was by a panel of academic staff and consultants from the academic departments of the faculty of medicine MAHSA University and clinical departments of the teaching hospital respectively. These staff were all involved in the clinical teaching programme of the faculty. Selected patients were then clerked by specialist trainee medical officers and case summaries prepared and held by the examinations unit of the faculty. During this process, all clinical year students were quarantined and not allowed access to the teaching hospital.

During the examination, all students were allowed to clerk the patient unobserved for one hour after drawing lots, followed by examination by a pair of pre-determined examiners for half an hour Each pair usually comprised an internal and external examiner paired by the examinations committee and allotted to a specific patient. Assessment and marks were based on an assessment rubric (Additional file - 1), with subjective comments on performance provided by the examiners. This assessment rubric was based on an existing rubric used in the faculty with the addition of an area to record subjective comments on performance. The use of this assessment rubric was approved by the Faculty Academic Board before the examination. All examiners were briefed on the assessment rubric by the chair of clinical examinations on the day of the examination. All examiners were advised to record their subjective comments to reflect the overall performance of the students. Examiners were also advised not to consider the numerical marks when recording their comments. To minimize observer bias both examiners were required to mark using the rubric provided and the final mark was consensual. The subjective comments were not used to decide a pass or fail status. Students would be deemed to have passed if they achieved a mark of 50%. For this study, the numerical scores and subjective comments were compared for concordance or otherwise and classified as “tally”, “does not tally” or “no comment”.

Inclusion criteria

Clinical long case results of all students were recorded for study with permission from the faculty.

Study variables

The clinical subjects, numerical marks given by examiners, their comments were written on the mark sheet and concordance between marks & comments were the study variables used.

Ethical committee approval

Ethical approval was obtained from the Institutional Research Board through the Faculty Research and Ethical Committee. Data was kept confidential and recorded as anonymous for analysis purpose.

Data management and statistical analysis

An assessment rubric for the clinical long case was used as a data collection tool (Additional file -1). The findings were recorded and analysed using Fisher's exact test since the assumptions for the Chi-square test were not met. The analysis was done for two separate groups: medicine & allied subjects and surgery & allied subjects. The point of significance was taken at 0.05. SPSS version 19.0 was used for data analysis.

Results

Mark sheets of all 105 students of that batch were reviewed and recorded. Overall concordance between numerical scores and subjective comments was seen in 32% of cases across both groups (medicine and allied subjects, surgery and allied subjects). Overall discordance was seen in 20%, and no comment was seen in 48% of cases across both groups.

When both groups were analysed separately, there was 37% concordance in the medicine and allied subjects group versus 28% concordance in the surgery and allied subjects group. Discordance was less in the latter group (15% versus 25%). (Fig. 1).

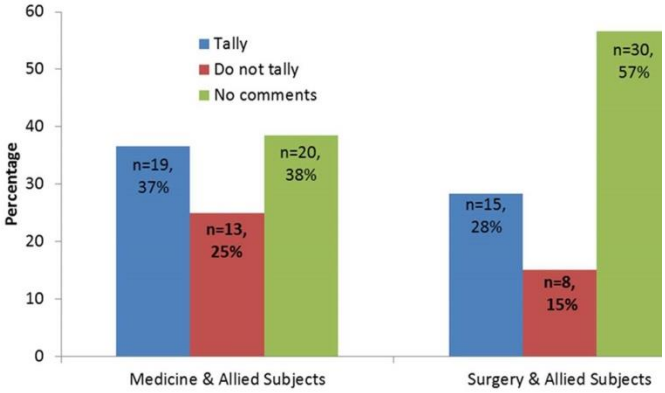

Figure 1: Concordance and Discordance of various subjects

The disparity was the lowest in the internal medicine and general surgery disciplines and highest in the psychiatry and paediatric disciplines, when each discipline's results were analysed separately. (Table 1 & 2). In about 48% of the assessment sheets, it was noted that there were no comments. Of these, all were clear passes except for two candidates, one of whom was deemed borderline and the other a clear fail.

Discussion

There was remarkable consistency seen in the numerical scores and subjective comments in internal medicine and general surgery (Table 3).

Table 1: Concordance of numerical scores to subjective comments

Disciplines	Tally	Do not tally	No comment	Total	Fisher's exact test p-value
	no. (%)	no. (%)	no. (%)	no.	
Medicine & Allied Subjects					
Medicine	11 (50)	1 (5)	10 (45)	22	0.004*
Paediatrics	8 (35)	7 (35)	5 (25)	20	
Psychiatry	0 (0)	5 (50)	5 (50)	10	
Surgery & Allied Subjects					
Surgery	7 (33)	1 (5)	13 (62)	21	0.161×
O&G	4 (18)	4 (18)	14 (64)	22	
Orthopaedics	4 (40)	3 (30)	3 (30)	10	
Total	34	21	50	105	

×p>0.05, †P<0.01

Table 2: Discordance of numerical scores to subjective comments

Discipline	No.	%	Subjective Comment	Scores given
Medicine	1	5	Poor knowledge	Borderline score
	2		Very poor knowledge	Borderline score
Paediatrics	4	35	Poor knowledge	Pass score
	1		Average knowledge	Excellent score
	1		Poor knowledge	Pass score
Psychiatry	2	50	Very poor knowledge	Borderline score
	2		Average knowledge	Strong Pass score
Surgery	1	5	Poor knowledge	Borderline score
O&G	3	18	Poor knowledge	Borderline score
	1		Good	Excellent score
Orthopaedics	3	30	Poor knowledge	Borderline score
Total	21	20		

In both these disciplines, the disparity was seen in only one candidate's scores. This is an indication of the clear consensus between examiners on the competencies that require testing at this level of examination and standardisation of marking schemes.

At the other end of the spectrum, there was no consistency seen in the assessment of students who were examined in the disciplines of psychiatry and paediatrics, which may be

indicative of a lack of clarity in the assessment of necessary competencies at an examination of this level, a lack of standardisation, or a reluctance to fail students. Furthermore, for 48% of candidates there were no comments recorded. As such, we found that subjective comments could not be reliably used in making decisions in borderline candidates in all disciplines.

Table 3: Numerical scores vs subjective comments (concordance)

Discipline	Marks scored	Subjective Comment
Medicine	44%	did not perform cardiac examination, did not know drug treatment of congestive cardiac failure
	56%	history – satisfactory, clinical exam – good, discussion – not satisfactory
Paediatrics	75%	excellence in all aspects of clinical clerking
Surgery	43%	poor history and physical examination. Not confident of findings, unable to discuss case
O&G	70%	good student
Orthopaedics	40%	poor history and examination technique, no clinical correlation

It has been found that this phenomenon of 'Failure to Fail' is a real problem for academics. The ramifications of such difficulty in high stakes professional examinations, such as a Final Professional Examination, are enormous. Among reasons suggested are a lack of knowledge on what to document in such cases, the emotional impact of failing on the academic and student and subsequent support required, difficulties with the remediation process, anxiety regarding consequences to the programme and reputation of the faculty and institution. [2, 3]

The long case has long been touted as a complete close to 'real life' patient encounter. [4, 5] This is because the student is required to collect and define information to help formulate a diagnosis and plan of management via history taking and physical examination of a patient in a manner very similar to real life practice.

The clinical component of the final professional examination is primarily a test of practical skills and applied knowledge and ideally should test at the level of 'does' in Miller's framework of assessment of competencies. [6] However, the traditional long case is hardly a reliable measure of competencies at this level; at its best, it will meet the requirements of a 'shows' in Miller's framework. This is because the pattern of running this format of examination entails an unobserved period of time when the student clerks the patient followed by a formal presentation in front of the examiner. This is followed by an oral examination that is frequently unstructured [7], which often does not take into consideration the difficulty level of the case.

Other problems in this type of examination include a lack of standardisation of patients, lack of standardisation in the competencies tested and expected competencies, and variability in scores given by different examiners. [8] There is also no assessment of the soft skills and clinical competencies required for interaction with the patient. [7] This may in part explain the variance in comments vis-à-vis numerical scores seen in our series.

This brings us to the question of how to overcome this discordance. An observed history taking and clinical examination with a standardised list of competencies to be tested would reduce inter and intraexaminer variability. [9, 10] A suitable method would be to use the Objective Structured Long Examination Record (OSLER). [7, 9] This sort of assessment will shorten the examination time, which is more in tune with the real clinical environment, allow demonstration of problem-solving skills, and allow a greater number of assessments with multiple examiners, resulting in a more valid and reliable examination result.

A global rating of performance, together with a numerical score based on a competency checklist, has been suggested as a method of improving the reliability of a clinical examination. These may be used together in deciding performance in a pass/fail situation. [1] A global rating is also an objective method of determining levels of achievement as it is standardised with an acceptable minimum level of achieved competencies to allow a pass in the examination.

Finally, training examiners in the use of a rubric or checklist is important to improve the ability of examiners to ensure passing students have the necessary competencies required of the assessment and reduce examiner variability, thus improving the validity and reliability of the examination.

Conclusion

It appears that this system of examination results in significant interexaminer variability depending on the type of long case presented to the student. This was in spite of case selection, use of standardized rubrics and the need for a consensual mark. This study also indicated a considerable variance in the concordance of marks and comments that appeared to be discipline specific.

This may be attributed to differences in the types of patients, differences in the examiner's expectations, lack of familiarity with the rubric, as well as a reluctance to fail even when subjective comments indicate low or suspicious competence. Using a global rating of competence with a structured rubric of required competencies and increasing the number of assessment encounters will result in greater concordance, validity and reliability.

A predetermined minimum competency level and score in global ratings and numerical scores required to pass will ensure greater consistency. Finally, training and repetitive use of a rubric or checklist will improve both the examiner

and checklist reliability to ensure fair and transparent assessment.

Limitation and future scope

It appears that concordance between subjective comments and marks obtained are high when there is consensus on required competencies for the examination is clear. The question remains as to why a lack of concordance occurs. It appears that this may be related to reluctance to fail a student in a high state examination. The reason for this reluctance may be varied. Implementing a more objective marking scheme (OSLER) would standardize the competencies that require testing. This may result in better decisions. It is clear that subjective comments have very limited use in deciding a pass/fail status. A global rating of performance would probably be more helpful in these situations.

Acknowledgement

The authors acknowledge the support of the Dean and participation of the examiners involved in the Final Professional Examination.

Authors' contribution

The design of this study involved all authors. Data curation was done by Dr AKKM. Analysis of data involved all authors. The final format of the article was decided upon by all authors and the final version was written by the corresponding author.

Competing interests

There is no financial, institutional or consultant conflict of interest for any author of this manuscript.

Publisher's Note

QIUP remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The publisher shall not be legally responsible for any types of loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

References

1. Mash B. Assessing clinical skills—standard setting in the objective structured clinical exam (OSCE). South African Family Practice. 2007;49(3):5-7. https://doi.org/10.1080/20786204.2007.10873520
2. Dudek NL, Marks MB, Regehr G. Failure to fail: the perspectives of clinical supervisors. Academic Medicine. 2005;80(10):S84-7.
3. Luhanga FL, Larocque S, MacEwan L, Yovita N, Danyluk P. Exploring the issue of failure to fail in professional education programs: A multidisciplinary study. Journal of University Teaching & Learning Practice. 2014;11(2):3.
4. van der Vleuten C. Making the best of the" long case". The Lancet. 1996;347(9003):704-5. https://doi.org/10.1016/S0140-6736(96)90069-0
5. Sood R, Singh T. Assessment in medical education: evolving perspectives and contemporary trends. The National medical journal of India. 2012;25(6):357-64.
6. Miller GE. The assessment of clinical skills/competence/performance. Academic medicine. 1990;65(9):S63-7. https://doi.org/10.1097/00001888-199009000-00045
7. Sood R. Long case examination-Can it be improved. J Indian Acad Clin Med. 2001;2(4):252-5.
8. Norcini JJ. The death of the long case?. BMJ. 2002;324(7334):408-9. https://doi.org/10.1136/bmj.324.7334.408
9. Gleeson F. AMEE medical education guide No. 9. Assessment of clinical competence using the Objective Structured Long Examination Record (OSLER). Medical Teacher. 1997 Jan 1;19(1):7-14. https://doi.org/10.3109/01421599709019339
10. Norman G. The long case versus objective structured clinical examinations: The long case is a bit better, if time is equal. BMJ: British Medical Journal. 2002;324(7340):748. https://doi.org/10.1136/bmj.324.7340.748

Additional material

(Additional file 1: Rubric Used in the Marking of the Clinical Long Case)

Click here for file