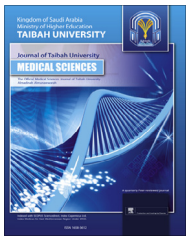
Journal of Taibah University Medical Sciences (2014) 9(2), 110–114

Taibah University

Journal of Taibah University Medical Sciences

www.sciencedirect.com



Educational Study

Difficulty index, discrimination index, sensitivity and specificity of long case and multiple choice questions to predict medical students' examination performance

Fahisham Taib, MRCPCH [a] and Muhamad Saiful Bahri Yusoff, PhD [b,*]

[a] *Department of Paediatric, Universiti Sains Malaysia, Kelantan, Malaysia*
[b] *Medical Education Department, Universiti Sains Malaysia, Kelantan, Malaysia*

Received 14 October 2013; revised 5 December 2013; accepted 7 December 2013

الملخص

أهداف البحث: إن تقييم كفاءة أطباء الغد يلعب دورا حيويا لتقديم نظرة ثاقبة لقدراتهم السريرية وإجمالي انجازاتهم. تبحث هذه الدراسة في مؤشر الصعوبة، ومؤشر التمييز، والمناطق الواقعة تحت منحنى خصائص المتلقي التشغيلية، وحساسية ونوعية مكونات التقييم المستخدمة في امتحان مقرر طب الأطفال في جامعة سينز ماليزيا.

طرق البحث: تمت دراسة استعادية بمراجعة سجلات أداء طلاب الطب أثناء الامتحانات. وكان المستهدفون طلاب السنة الرابعة في عام 2012م (ن = 210) و 2013م (ن = 177) الذين جلسوا لامتحانات نهاية مقرر طب الأطفال بعد اكمالهم 6 أسابيع من التدريب وقد تضمن الامتحان حالة سريرية طويلة وأسئلة متعددة الخيارات.

النتائج: كان مؤشر الصعوبة للأسئلة متعددة الخيارات يتراوح بين 0.67-0.79 ويعتبر مستوى هذا المؤشر مثاليا. بينما تراوح مؤشر الصعوبة للحالة السريرية الطويلة بين 0.89-0.91 ويعتبر مستوى هذا المؤشر أقل مثالية. كما أظهرت الدراسة ارتفاع مؤشر التمييز للأسئلة متعددة الخيارات (0.58-0.76)، مقارنة بامتحان الحالات السريرية الطويلة (0.20-0.23)، مما يدل على أن الأسئلة متعددة الخيارات أكثر قدرة من امتحان الحالة السريرية الطويلة على التمييز بين الطلاب الجيدين من غيرهم.

الاستنتاجات: هناك دلائل كثيرة على قدرة الأسئلة متعددة الخيارات على التفريق بين الطلاب الجيدين من غيرهم مع وجود مستوى صعوبة أفضل من امتحان الحالة السريرية الطويلة. كما يبدو أن المقاييس النفسية للأسئلة متعددة الخيارات جيدة لتغطيتها معلومات طبية واسعة في وقت أقصر، بينما المقاييس النفسية لامتحان الحالات السريرية الطويلة ضعيفة بسبب ذاتية التقييم.

الكلمات المفتاحية: الأسئلة متعددة الخيارات; الحالات السريرية الطويلة; التقييم; مؤشر الصعوبة; مؤشر التمييز

Abstract

Objectives: The competency assessment of tomorrow's doctors plays a vital role to offer insight into their clinical abilities and overall achievement. This study explores difficult index, discrimination index, areas under ROC curve, sensitivity and specificity of assessment components employed in the pediatric examination in Universiti Sains Malaysia (USM).

Methods: A retrospective record review of medical undergraduates' examination performance was done. The target population were fourth-year medical students in 2012

* Corresponding address: Medical Education Department, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian, 16150 Kota Bharu, Kelantan, Malaysia. Tel.: +60 97676550; fax: +60 97653370.
E-mail: msaiful_bahri@usm.my (M.S.B. Yusoff)

Peer review under responsibility of Taibah University.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jtumed.2013.12.002

(n = 210) and 2013 (n = 177) academic year that sat for the pediatric end posting examinations after completing a 6-week rotation. Each of the examinations comprised of MCQ and Long Case.

Results: The difficulty index of MCQ ranged from 0.67 to 0.79, which is considered as optimal level. The difficulty index for Long Case ranged from 0.89 to 0.91, which is considered as less optimal level. The MCQ demonstrated higher discrimination index (0.58–0.76) than the long case (0.20–0.23), suggesting the MCQ was better able to discriminate poor and good students than the long case.

Conclusion: MCQ has more evidence to support its discriminant validity and optimal difficulty level than the long case for both cohorts of medical students. The MCQ has good psychometric credentials which may results of the broad sampling of knowledge over short duration of time, while the long case seems to have poor psychometric credentials which may results of the assessment subjectivity.

Keywords: Assessment; Difficulty index; Discrimination index; Long case; Multiple choice question

Introduction

Competency assessment is the backbone in the clinical education of tomorrow's doctors. The method assessment look likes to shape students' learning approaches and performances.[1–4] Therefore, proper design of assessment formats will surely drive ways of students approach to learning. Inappropriate design of assessment formats may lead to unwanted outcomes of competencies and types of patient care.[3] The best assessment plays a vital role to offer insight in students' clinical abilities and overall achievement.[1–3] Epstein et al. (2007) described competence as "a habit of lifelong learning, contextual that reflecting person ability to performing tasks and developmental in nature, where it is a result of a well planned practice and reflection on own experience". These characteristics add to the realism of an assessment.

For the past few decades, many medical education programs and licensing authorities either at undergraduate level or postgraduate level have allocated tremendous efforts to ensure the authenticity of assessments and competency of trainees.[1,5] Every assessment format has its advantages and disadvantages depending on the assessment design. The best assessment method must meet five criteria which include reliability, validity, acceptability, feasibility and educational impacts on learning and practice.[6] Miller classified assessment methods into four categories which include knows (i.e. assessing knowledge), knows how (i.e. assessing ability to apply knowledge within its context), show how (i.e. assessing trainees' performance in simulated environment) and does (i.e. assessing trainees' performance in actual environment).[7] The 'does' component is considered as the most difficult area to be examined authentically.[7]

Validity is commonly defined as the degree of a measurement really gauges characteristics it is intended to assess.[6–9] Sources of evidence to support validity can be gathered in the forms of content, response process, internal structure, relations to other variables and consequences.[8] Content validity refers to the extent of intended learning outcomes covered by an assessment through a proper content blueprint.[6,7] Conversely, it is achieved when test items are adequately covering expected learning outcomes of a course; this is known as content relevant.[7] Validity-related to response process is achieved when there are substantial relationships between the test item components and the subjects' thought process.[8] Internal structure is related to the correlation between test items of an assessment tool.[8] Validity-related to relations to other variables is achieved when it correlates with other assessment tools that measure similar characteristics.[6–8] Validity-related to consequences is signified when test items of an assessment predict educational variables such as quality of patient care and doctor–patient relationships.[6–8] To improve the validity of an assessment, Epstein (2007) recommended four actions which include (1) clear expectation of an assessment, (2) clear learning outcomes to be measured, (3) familiar with the advantages and disadvantages of an assessment too, and (4) continuous evaluation and monitoring of assessment quality to avoid the unwanted effects.

Considering these facts, this article explores the source of validity evidence in assessment components employed by the pediatric examination for the fourth year undergraduate medical students in Universiti Sains Malaysia (USM) that include difficult index, discrimination index, areas under ROC curve, sensitivity and specificity. It is hoped that this will be able to provide insight about quality of the assessment.

Materials and Methods

Study setting

A retrospective record review was done on medical undergraduates' examination performance in the Department of Pediatric USM. Approval to conduct this study was obtained from the Department and School of Medical Sciences, USM.

The target population comprised of fourth-year undergraduate medical students in USM for the 2012 (n = 210) and 2013 (n = 177) academic year that undertaken the end of posting exams (MCQ and Long Case) after completing a 6-week pediatric rotation. Each end of posting exams comprised of MCQ and Long Case as the assessment method for each group that was attached to the pediatric department. During that period, 'pediatric apprenticeship model' (Taib, 2013[17]) has been used extensively since 2009 as part revitalized pediatric program. This is a self directed learning model integrating clinical and problem solving learning. Students are required to learn by following the senior doctors during the clinical attachment. The learning and opportunistic discussion regarding specific pediatric cases are discussed during the clinical rotation. The students will have to plan and decide for their learning issues by self inquiry clinical questions and proactively search the answer through research and discussion. During that process, students are required to complete a logbook, presentations and case write up. Students are exposed to various clinical encounters which would essentially help them to improve both core knowledge to answer MCQ and clinical skills for the Long Case.

Assessment components and test formats of the pediatric examination

Assessment methodology for the department consists of 3 important areas. Attitude is assessed by routine attendance and report from clinical supervisor. Despite its subjective nature of assessment, the decision for passing end of posting must also be guided on satisfactory attendance and completion of the logbook requirement.

The MCQ was organized at the end of the posting and it consists of 30 questions which were selected randomly. These are standardized and vetted questions are selected from past years' final professional exam questions. The pediatric coordinator will randomly select the question based on the availability of recent vetted bank questions and according to systems and its suitability. Standardization of questions is achieved with standard mixture of pass year and newly vetted questions. The level of difficulty is tailored according to undergraduate MCQ requirement and each group will sit for different set of questions. This move is to avoid repetitive questions and plagiarism to the next group of students.

Clinical Long Case usually was set up by individual examiner in the final 2 weeks of pediatric rotation. Standardization of cases was not made but common 'bread and butter' cases were taken as index case for year 4 students. Cases allocated usually are fresh and newly admitted patients into pediatrics ward. The cases are also guided by what are required in logbook and important clinical aspects for discussion. The context of assessment in Long Case depends on the presentation skills, clinical skills and discussion aspect. Student will also get examiner in a randomly fashion and the examiner should not be their personal supervisor.

Statistical analysis

All analyses were performed by using SPSS version 20.

Difficulty index is defined as the percentage of those candidates recording either a true or false response for a particular branch in a multiple true–false response MCQ who gave the correct response.[10,11] The optimal range is 20–80%; a low index may mean that students are attempting the item but are getting it wrong and a too high index may mean that regardless of poor or good students are able to get it correct. In this study, the difficulty index was determined by the percentage of students who passed in the MCQ and Long Case examinations.

Discrimination index is a measure, of how the 'good' students are doing versus the 'poor' students on a particular question. Knowing this, we expect the value of the discrimination index to range between 1 (all 'good' students correct versus no 'poor' students correct by the former method or the maximum value for a positive correlation by the latter method) to −1.[11,12] Discrimination index of 0.40 and up is considered as very good items, 0.30–0.39 is reasonably good, 0.20–0.29 is marginal items (i.e. subject to improvement), and 0.19 or less is poor items (i.e. to be rejected or improved by revision).[11,12] The discrimination index was calculated by the SPPS 20.

The sensitivity, specificity, and area under receiver operating characteristics (ROC) curve of MCQ and Long Case were performed by ROC analysis to signify their ability to predict and discriminate poor and good performances. The sensitivity, specificity and area under ROC curve values more than 0.70 were considered as having an acceptable predictive and discriminative value, while more than 0.8 is considered as very good level.[13]

Results

387 4th year medical students' MCQ and Long Case marks were obtained from the pediatric department; 210 4th year medical students in 2012 cohort and 177 4th year medical students in 2013. Permission was sought from the department prior to data collection.

In general, for both cohorts, the difficulty index of MCQ ranged from 0.67 to 0.79, which is considered as optimal level of difficulty[10,11] (Table 1). Conversely, the difficulty index for Long Case ranged from 0.89 to 0.91, which is considered as less optimal level of difficulty (i.e. might be too easy)[10,11] (Table 1). The MCQ in 2013 cohort demonstrated higher discrimination index than the MCQ in 2012 cohort, suggesting the 2013 MCQ was better able to discriminate poor and good students than the 2012 MCQ.[11,12] In contrast, the Long Case has marginal discrimination index, suggesting that they are not really able to discriminate poor and good students.[11,12]

The ROC analysis (Table 2 and Fig. 1) showed that MCQ consistently had very good level of discriminative ability as the value more than 0.8.[13] On top of that its sensitivity and specificity values were more than 80%, indicating good level of ability to predict and discriminate passed and failed students.[13] In other hand, Long Case demonstrated poor discriminative ability as the ROC values were less than 0.7.[13]

Discussion

Our results showed that MCQ was at optimal level of difficulty and able to discriminate performance of poor and good

Table 1: The difficulty index and discrimination index of MCQ and Long Case in the end-posting pediatric examination of two cohorts of 4th year medical students.

Student cohort	Assessment tool	Difficulty index	Discrimination index
2012	MCQ	0.79	0.56
	Long Case	0.91	0.23
2013	MCQ	0.67	0.78
	Long Case	0.89	0.20

Table 2: The area under ROC curve, sensitivity and specificity values of MCQ and Long Case in order to predict pass–fail outcomes in the end-posting pediatric examination for two cohorts of 4th year medical students.

Student cohort	Assessment tool	ROC (CI 95%)	Sensitivity (%)	Specificity (%)
2012	MCQ	0.88 (0.81, 0.96)	88.0	88.5
	Long Case	0.67 (0.54, 0.80)	95.1	38.5
2013	MCQ	0.90 (0.86, 0.95)	83.6	97.3
	Long Case	0.58 (0.47, 0.69)	92.1	24.3

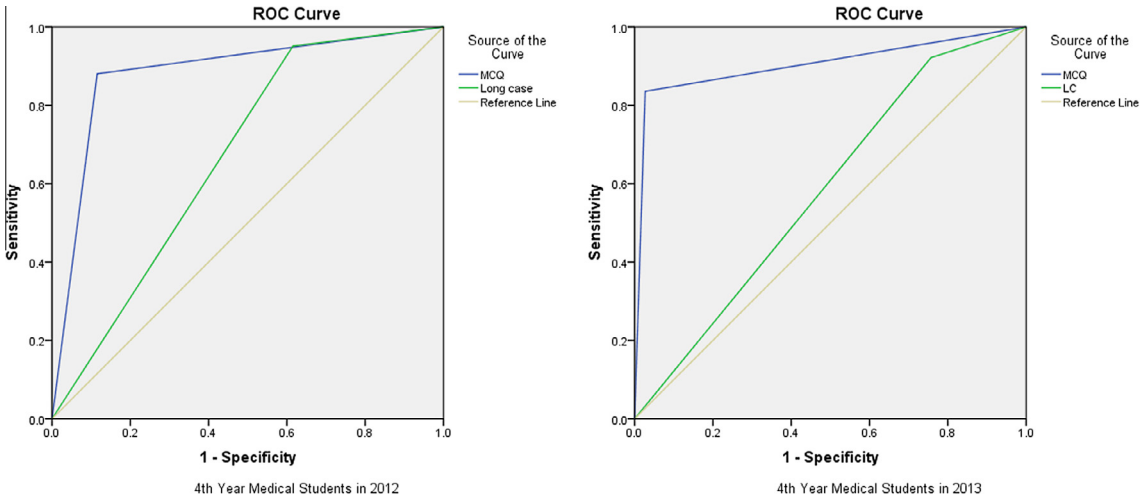


Figure 1: Comparison on sensitivity and specificity of MCQ and Long Case to predict pass–fail outcomes of the end-posting pediatric examination of two student cohorts.

students. This finding may be due to the broad sampling of knowledge content over short duration of time as compared to the long case.[6] Interestingly, it seems that 2013 MCQ was superior to 2012 MCQ in terms of discriminative ability. One possible reason is due to the coordinator of the examination has administered a new set of vetted MCQ to the 2013 cohort. The selection of questions in MCQ usually is done in random arrangement to cover as much clinical learning. The MCQ marking is negatively marked and the students are advised to answer only confirmed known correct answer.

Our results suggested that long case examination was easy and unable to discriminate performance of poor and good students. One of the possible reasons is due to the variability of examiners' rating judgment because of the assessment subjectivity.[14–16] Considering that year 4 students were exposed with introductory phase of pediatrics, skills obtained during the posting may vary depending on the clinical environment and experience. Long cases for both cohort students (2012 and 2013) have low specificity. Performance of the students during the long case depended on the case given during the end of posting exam. Students may have obtained adequate clinical skills during the apprenticeship in the clinical setting.[17] The case given may be well discussed before hand during seminars or clinical presentation. The rater judgment may be much lenient in view of introductory phase of pediatric learning. There are other factors which may influence rater judgment. There are potential humanistic values during the interaction and communication between the students, the patents and the examiner; the knowledge of the strength and weakness of the student prior to exam of which supervisor and examiner may have communicated before hands; the expectation and low bench marking in view of consideration that year 4 is considered as introductory rather than advanced knowledge in pediatrics; maturity of students during their dealing and history taking; potential hiccups during the exam for example the child starts crying and in order to divert the attention of anxious candidate, the examiner discuss superficially about the patient. These potential factors, despite having its draw back in differentiating between good and bad students, give a balance scope of assessment by recognizing altruistic values in assessment where there is none in MCQ.

The potential sources of rating errors are examiners, rating forms or scales, rating items or tasks, and rating objects or subjects.[18] Among them, examiner variability significantly contributing to the rating errors.[18] Examiners' judgments on performance of students are known to be susceptible to generosity bias (i.e., tendency to give positive remarks) and frequently fail to detect real discrepancy that compromises the validity of rating judgment.[14–16] Several ways were recommended to address issues related to validity of rating judgments. First, examiners are trained to observe the competencies to be assessed thus their expert judgments are standardized.[14,18] Second, examiners are made known to the rating formats, so that they get a clear view on ways for effectively rating examinees' performance.[14,18] Third, adopting triangulation of multiple assessment for judging competencies might lessen the judgment errors, thus may improve the validity of examiners' judgment.[14,18] For example, decisions on examinees' clinical competencies are made based on multiple long cases and other assessment tools such as short cases. Fourth, simplify the rating task might improve the validity of examiners' judgment due to reducing unnecessary cognitive load during rating process.[14,18] Fifth, proper assessment blueprinting might provide a guide to calibrate examiners' expectation thus it may lessen judgment variability between examiners. In addition, maximizing inter-examiner reliability could reduce the judgment errors caused by examiners.[14,18]

Several limitations need to be highlighted for interpretation and future research. First, the researcher conducted this study on one educational setting, which limited generalizability of the results. Therefore any effort to infer this finding to other educational settings must be done within context. Second, several variables that might influence the study outcomes such as examiner characteristics, previous academic performance of students and psychological health status of students were not controlled during the analysis therefore accuracy of the results might be questionable. Lastly, the overall performance of end of posting as the reference point to calculate discrimination index, difficulty index, area under ROC, sensitivity and sensitivity might not be the best standard, thus the accuracy of the results might be compromised.

Conclusion

Our study found that MCQ has more evidence to support its validity to discriminate performance of poor and good students. The long case seems to have less evidence to support its validity to discriminate performance of poor and good students, which may results of examiner variability. However, despite these differences measures should be taken to improve the validity of long cases which were discussed in this article.

Conflict of interest

The authors have no conflict of interest.

References

1. Epstein RM. Assessment in medical education. **N Engl J Med 2007**; 356(4): 387–396.
2. Barman A. Critiques on the objective structured clinical examination. **Ann Acad Med Singapore 2005**; 34(8): 478–482.
3. Gulijers JTM, Bastiaens TJ, Kirshner PA. A five-dimensional framework for authentic assessment. **Educ Technol Res Dev 2004**; 52(3): 67–86.
4. Karim K, Edwards R, Dogra N, et al. A survey of the teaching and assessment of undergraduate psychiatry in the medical schools of the United Kingdom and Ireland. **Med Teach 2009**; 31(11): 1024–1029.
5. Newble D. Assessing clinical competence at the undergraduate level. **Med Educ 1992**; 26(6): 504–511.
6. van der Vleuten C. Validity of final examinations in undergraduate medical training. **BMJ 2000**; 321(7270): 1217–1219.
7. Miller G. The assessment of clinical skills/competence/performance. **Acad Med 1990**; 65(9): S63–S67.
8. Cook DA, Beckman T. Current concepts in validity and reliability for psychometric instrument: theory and application. **Am J Med 2006**; 119(166): e7–e16.
9. Streiner D, Norman G. *Health measurement scales: a practical guide to their development and use*. 4th ed. New York: Oxford University Press; 2008.
10. Dixon R. Evaluating and improving multiple choice papers: true–false questions in public health medicine. **Med Educ 1994**; 28(5): 400–408.
11. Rahim AFbA. *What those number mean?* 1st ed. Kubang Kerian: KKMED; 2010, Downloadable at: http://www.medic.usm.my/dme/images/stories/staff/KKMED/2010/item_analysis_guide.pdf.
12. Ebel RL, Frisbie DA. *Essentials of educational measurement*. 5th ed. Englewood Cliffs, New Jersey: Prentice-Hall Inc; 1991.
13. Thomas GT. Interpreting diagnostic test; 2009. http://gim.unmc.edu/dxtests/Default.htm [accessed 2009].
14. Albanese MA. Challenges in using rater judgments in medical education. **J Eval Clin Pract 2000**; 6(3): 305–319.
15. Yusoff MSB. Discrepancy-agreement grading provides feedback on rater judgments. **Med Educ 2012**; 46(11): 1122.
16. Yusoff MSB, Abdul Rahim AF. The discrepancy-agreement grade (DAG): a novel grading system to provide feedback on rater judgments. **Educ Med J 2012**; 4(2): e100–e104.
17. Taib F, Mat Zin MR, Ab Majid N, Yusoff MSB, Hans VR. Apprenticeship teaching in paediatrics: students perspective for future improvement. **Int Med J 2013**; 20(1): 1–3.
18. Downing SM. Threats to the validity of clinical teaching assessments: What about rater error? **Med Educ 2005**; 39(4): 353–355.