Contribution of Examiner Subjectivity and Patient Heterogeneity on Long Case Examination

Nitipatana Chierakul MD*,**, Somwang Danchaivijitr MD*,**,
Paka Kontee BBA*, Chana Naruman MSc**

** Subcommittee for Training and Examination, The Royal College of Physicians of Thailand, Bangkok, Thailand*
*** Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand*

Objective: *To evaluate the effect of examiner subjectivity and heterogeneity among the cases on scores from the Royal College of Physicians (RCPT) long case examination.*
Material and Method: *Data from internal medicine candidates who performed clinical part of RCPT board certifying examination in academic year 2008 were collected. For each candidate, scores from pair of examiners for each of the long case was stratified based on disease category according to the course syllabus into 3 groups; very common, common and uncommon diseases. The scores also categorized according to difficulty level subjectively and rated by the examiners into 3 levels; easy, moderate and difficult. Mean scores in each group of encounters were compared using ANOVA.*
Results: *There were 21 examination centers involved with 1,840 number of encounters by 232 candidates. Among 437 patients that have been used for the long case, common scenarios (27.6% of the total) were cirrhosis, hyperthyroidism, cerebral thrombosis, bronchogenic carcinoma, rheumatic heart disease and thalassemia. Mean and SD of scores from the very common, common and uncommon diseases were 75.5 $\pm$ 11.6, 75.6 $\pm$ 10.6 and 74.7 $\pm$ 11.3 respectively, with no statistical significant difference between the groups. Mean and SD of scores from the easy, moderate and difficult cases were 76.1 $\pm$ 10.5, 74.8 $\pm$ 11.0, 75.5 $\pm$ 10.9 respectively. The moderate group has the lowest score with a statistical significant difference from other groups ($p = 0.042$).*
Conclusion: *In current RCPT long case examination, difficulty of the case appears to contribute to variation in scores derived from the examiners. Measures for score adjustment and examiner calibration should be implemented in the future.*

Keywords: *Long case, Patient heterogeneity, Examiner heterogeneity*

J Med Assoc Thai 2012; 95 (Suppl. 2): S83-S86
Full text. e-Journal: *http://www.jmat.mat.or.th*

Interaction between the candidate and the real patient in long case examination has been used for a long time in the clinical part of the Thai Board of Internal Medicine certifying examination. The candidate is expected to obtain relevant data from the task of history taking and physical examination, generate a hypothesis, and finally express a management plan concerning the context of a specific patient encountered. Whether this kind of examination is reliable or valid in terms of examiner subjectivity, heterogeneity among the cases, and aspect of competence assessed have been raised among medical educators[1-3].

The authors have demonstrated the acceptable inter-rater reliability among a pair of long case examiners in the high stake board certifying examination held by the Royal College of Physicians of Thailand (RCPT)[4]. In the present study, the authors aim to evaluate the effect of examiner subjectivity and heterogeneity among the cases on scores from long case examination.

Correspondence to:
Chierakul N, Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand.
Phone: 0-2419-7757
E-mail: siade@mahidol.ac.th

Material and Method

The RCPT long case examination

There are 2 occasions of RCPT clinical examination during the third year of internal medicine resident, each candidate must perform 4 encounters of long case examination which contribute 58% of the total score for clinical examination. For each examination center, there are 6 or 12 candidates randomly allocated to perform the test under 12 or 24 examiners approved by the RCPT.

During a long case, candidates encounter the patient under direct observation of pair examiners during the period of 30 minutes for history taking and carrying

out physical examination. Then they present their findings, provisional diagnosis and management plan to the examiners. Some questions and direct interpretation of clinical specimens are also be added by the examiners. Informing and educating the patient are the last task for the total 70-minute examination. After the candidate finishes each long case, examiners independently note, on a 5-level scale, for each section of the structured marking form for score transfer.

Before the assessment, both examiners spend 30 minutes with the patient to confirm or adjust clinical notes prepared by examination center. Examiners also stratify each of the long cases based on disease category, according to the course syllabus, into very common, common, and uncommon diseases. Difficulty level is also subjectively rated by the examiners into easy, moderate, and difficult for each patient.

Data collection

Data from internal medicine candidates who performed clinical part of RCPT board certifying examination in academic year 2008 were collected. For each candidate, scores from pair of examiners were divided into 3 groups according to the disease category and the difficulty level; they are presented in the form of mean and standard deviation (SD). Comparison of score between each group was performed using 1-way ANOVA and LSD method for post-hoc multiple comparisons. All statistical analyses were carried out using statistical software SPSS version 13.0 (SPSS Inc, Chicago, USA).

Results

There were 21 examination centers with 464 examiners and 232 candidates in 2008 RCPT clinical examination. Among 437 patients that have been used for the long case, the common scenarios (27.6% of the total) are shown in Table 1. Two-third of the cases was categorized as common disease and the mean with SD of scores from each disease category are demonstrated in Table 2. There were no statistical significant difference between scores from each disease category. About half of the cases were subjectively rated the difficulty as moderate and the score from this level was significant lower than the easy and the difficult groups, p = 0.042 (Table 3).

Discussion

Competence assessment evaluates the ability to do during training whilst performance assessment evaluates the actual doing in clinical practice[5]. Although enormous variations on the long case are the matter of concern, RCPT still put a long case as a major section of clinical part of board certifying examination. During the long case, examiners are required to see the patient immediately before the candidate for reducing the bias from examiner's own specialty and information offered by examination center. No significant difference among pairs of examiners from different specialties and institutions has been demonstrated in the authors' previous study[4].

In term of case heterogeneity according to

Table 1. Common scenarios that have been use in RCPT long case examination 2008

Scenario	%
Cirrhosis	3
Hyperthyroidism	3
Cerebral thrombosis	2.3
Bronchogenic carcinoma	2.3
Rheumatic heart disease	2.3
Thalassemia	2.3
Infective endocarditis	2.1
Pulmonary hypertension	2.1
Systemic lupus erythrematosus	1.8
Multiple myeloma	1.6
Hyperaldosteronism	1.6
Diabetes mellitus	1.6
Spinal cord compression	1.6

Table 2. Mean and standard deviation of scores from different disease category

Category	Number	Scores
Very common	376	75.5 $\pm$ 11.6
Common	1257	75.6 $\pm$ 10.6
Uncommon	207	74.7 $\pm$ 11.3

Table 3. Mean and standard deviation of scores from different difficulty level

Level	Number	Scores
Easy	587	76.1 $\pm$ 10.5
Moderate	855	74.8 $\pm$ 11.0
Difficult	398	75.5 $\pm$ 10.9

p = 0.034 between moderate and easy group
p = 0.043 between moderate and difficult group

the course syllabus, no significant effect on scores was demonstrated in the present study. Examiner may reduce their threshold for giving the score if the candidate encounter with patient harbored uncommon disease. Because each candidate has a chance to perform the examination in 4 different patients at 2 different examination centers, so the random chance to depend on the "luck of the draw" will be diluted. Increase the number of encounter of a long case up to 4 encounters or more under direct observation and a structured manner have also been shown to improve the reliability of this kind of competence assessment[6-10].

Examiner subjectivity has some significant effect on scores in the group with moderate difficulty which comprised about half of the total patients. The examiners may adjust their rating behavior by their own perception for the difficulty level of patients. A statistical model that removes difference behavior among "Hawk" and "Dove" examiners may improve the reliability of scores.

Current clinical examination, especially long case, is labor intensive for the examination centers, requires strenuous effort from the examiners and creates high expenses both for the RCPT as an organizer and also the candidates. Practice long cases undertaken in the workplace (formative evaluation) has been shown to be less reliable to those undertaken during examination conditions (summative evaluation)[11]. Multimodalities in-training evaluation has been suggested for its potential to assess all round clinical competence[12]. Further development should be considered for the substitution of some or major parts of the RCPT clinical examination with the proposed conventional and novel tools.

Conclusion

For the current RCPT long case as a main fraction of clinical part of board certifying examination, variation in scores given by the examiners resulted from the case difficulty. Measures for score adjustment and examiner calibration should be implemented in the future to create more fairness in such a high stake examination.

Potential conflicts of interest

None.

References

1. Norcini JJ. The death of the long case? BMJ 2002; 324: 408-9.
2. Wass V, van der Vleuten C. The long case. Med Educ 2004; 38: 1176-80.
3. Ponnamperuma GG, Karunathilake IM, McAleer S, Davis MH. The long case and its modifications: a literature review. Med Educ 2009; 43: 936-41.
4. Chierakul N, Danchaivijitr S, Kontee P, Naruman C. Reliability and validity of long case and short case in internal medicine board certification examination. J Med Assoc Thai 2010; 93: 424-8.
5. Boursicot K, Etheridge L, Setna Z, Sturrock A, Ker J, Smee S, et al. Performance in assessment: consensus statement and recommendations from the Ottawa conference. Med Teach 2011; 33: 370-83.
6. Gleeson F. Assessment of clinical competence using the Objective Structured Long Examination Record (OSLER). Med Teach 1997; 19: 7-14.
7. Wass V, Jones R, van der Vleuten C. Standardized or real patients to test clinical competence? The long case revisited. Med Educ 2001; 35: 321-5.
8. Wass V, Jolly B. Does observation add to the validity of the long case? Med Educ 2001; 35: 729-34.
9. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. Ann Intern Med 2004; 140: 874-81.
10. Wilkinson TJ, Campbell PJ, Judd SJ. Reliability of the long case. Med Educ 2008; 42: 887-93.
11. Wilkinson TJ, D'Orsogna LJ, Nair BR, Judd SJ, Frampton CM. The reliability of long and short cases undertaken as practice for a summative examination. Intern Med J 2010; 40: 581-6.
12. Holmboe ES, Hawkins RE. Methods for evaluating the clinical competence of residents in internal medicine: a review. Ann Intern Med 1998; 129: 42-8.

ผลของความแปรปรวนจากกรรมการผู้คุมสอบและความหลากหลายของผู้ป่วยต่อการสอบรายยาว

นิธิพัฒน์ เจียรกุล, สมหวัง ด่านชัยวิจิตร, ผกา คณฑี, ชนะ นฤมาน

วัตถุประสงค์: *เพื่อประเมินผลของอัตวิสัยของกรรมการผู้คุมสอบและความหลากหลายของประเภทผู้ป่วยที่นำมาสอบต่อคะแนนในการสอบรายยาวเพื่อวุฒิบัตรของราชวิทยาลัยอายุรแพทย์แห่งประเทศไทย*

วัสดุและวิธีการ: *รวบรวมข้อมูลของผู้เข้าสอบภาคปฏิบัติเพื่อวุฒิบัตรของราชวิทยาลัยอายุรแพทย์แห่งประเทศไทยในปีการศึกษา พ.ศ. 2551 จัดแบ่งคะแนนที่ผู้เข้าสอบแต่ละคนได้จากกรรมการคู่สอบแต่ละคนตามประเภทของผู้ป่วยที่ระบุในหลักสูตรเป็น 3 กลุ่ม คือ โรคที่พบบ่อย พบได้ทั่วไป และ พบน้อย พร้อมกับจัดแบ่งคะแนนตามระดับความยากที่กรรมการใช้อัตวิสัยกำหนดไว้เป็น 3 ระดับ คือ ง่าย ปานกลางและยาก ทำการเปรียบเทียบคะแนนเฉลี่ยระหว่างกลุ่มย่อยต่างๆ โดย ANOVA*

ผลการศึกษา: *มีสนามสอบทั้งสิ้น 21 แห่ง โดยมีการให้คะแนนทั้งหมด 1,840 ครั้งสำหรับผู้เข้าสอบ 232 คน โดยมีการใช้ผู้ป่วย 437 ราย โรคส่วนใหญ่ที่นำมาสอบ (ร้อยละ 27.6 ของทั้งหมด) คือตับแข็ง ไทรอยด์เป็นพิษ สมองขาดเลือด มะเร็งปอด โรคหัวใจรูมาติก และธาลัสซีเมีย ค่าเฉลี่ยของคะแนนจากประเภทของผู้ป่วยด้วยโรคที่พบบ่อย พบได้ทั่วไป และพบน้อย คือ 75.5 $\pm$ 11.6, 75.6 $\pm$ 10.6 และ 74.7 $\pm$ 11.3 คะแนนตามลำดับ ซึ่งไม่แตกต่างอย่างมีนัยสำคัญทางสถิติ ค่าเฉลี่ยของคะแนนจากระดับความยากของผู้ป่วยที่ง่าย ปานกลาง และยาก คือ 76.1 $\pm$ 10.5, 74.8 $\pm$ 11.0 และ 75.5 $\pm$ 10.9 คะแนนตามลำดับ กลุ่มที่ถูกจัดเป็นระดับความยากปานกลาง มีคะแนนเฉลี่ยต่ำกว่ากลุ่มอื่นอย่างมีนัยสำคัญทางสถิติ ($p = 0.042$)*

สรุป: *การสอบรายยาวเพื่อวุฒิบัตรของราชวิทยาลัยอายุรแพทย์แห่งประเทศไทยในปัจจุบันความยากของผู้ป่วยที่ถูกนำมาใช้สอบมีผลต่อคะแนนที่ได้รับจากกรรมการผู้คุมสอบ ในอนาคตควรมีมาตรการเพื่อปรับชดเชยคะแนนและปรับมาตรฐานการให้คะแนนของกรรมการ*