Reliability and Validity of Long Case and Short Case in Internal Medicine Board Certification 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*
***Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand*

Objective: *To be certified for the Thai Board of Internal Medicine, each candidate must pass both written and clinical examinations performed in different academic years. The present study aimed to assess the reliability and validity of the long case and short case which contribute major fractions in the clinical part of board certifying examination.*
Material and Method: *Data from 585 internal medicine residents entering a first-attempt clinical part in board certifying examination during the academic year 2005-2007 were collected. Inter-rater reliability and construct validity of the long case and short case were then examined.*
Results: *Good to excellent intraclass correlation (ICC) of scores from different examiners was demonstrated (ICC between 0.71 and 0.97) and the variation ranged from 15.3 to 27.3%. For different occasions of examination, class normalized gain was between -0.7 and -9.0% and negative individual normalized gain was observed in 45.6% to 48.2% of the candidates.*
Conclusion: *Acceptable inter-rater reliability was demonstrated in long case and short case in clinical examination for the Thai Board of Internal Medicine. But construct validity for this type of clinical assessment was not established.*

Keywords: *Internal medicine, Board certifying examination, Long case, Short case*

J Med Assoc Thai 2010; 93 (4): 424-8
Full text. e-Journal: *http://www.mat.or.th/journal*

The Thai Board of Internal Medicine certifying examination comprises of written and clinical parts. The candidates must pass both parts for achieving a diploma. The long case and short case have been used for a long time in clinical examination because they closely resemble important tasks in daily practice. Numerous arguments concerning the reliability and validity of this type of examination have been raised[1-3]. The examiner subjectivity, heterogeneity among the cases, and aspect of competence assessed are the main areas of discussion.

The authors' earlier study demonstrated the modest correlation between scores from written and clinical parts of board certifying examination held by the Royal College of Physicians of Thailand (RCPT)[4]. In the present study, the authors aimed to evaluate the reliability and validity of long case and short case which contribute the major sections of clinical 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 clinical examination

Traditionally, clinical examination was held by the RCPT at the end of the third-year training. In 2005, clinical examination was split into two occasions, at the middle and the end of the third-year training. The mid-year examination which comprised two long cases (each with 15% of the total scores), and the end-of-year examination consisted of one long case (20%), six short cases (30%), and ten laboratory stations (20%). The structure of examination was changed in 2007, both mid-year and end-of-year examinations had the same structure, two long cases, three short cases, and five laboratory stations. But the scores during the mid-year examination were only one-third of the total for clinical examination.

For long cases and short cases, there were 2 examiners for each candidate, one from the examination hospital and the other from a different hospital which has an internal medicine training program. The examiners are appointed by the RCPT, they must be aged between 35-65 years, have internal medicine or

related board certification, and worked as an internist for more than 10 years (including years in training). Instruction for the expected role and structured marking form is distributed to each examiner by the representative of each training center.

The RCPT long case is a 75-minute observed encounter between candidate and patient that focuses on six competences: history taking; physical examination; proper investigation; synthesizing the findings; developing a management plan; and informing and educating the patient. Two examiners spend 20 minutes with the patient before the assessment to confirm or adjust clinical notes prepared by examination hospital. 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.

During a short case, each candidate spends 5 minutes with the patient to perform the proper process of physical examination of a focus system and to detect signs, then a further 3 minutes to formulate a clinical or differential diagnosis. Two examiners observe the encounter and note on a 5-level scale of the structured marking form prepared by the RCPT for process of physical examination and an open-end score for the outcome of findings. In some cases the communication skill and professionalism are assessed instead of physical examination.

Data collection

Between the academic years 2005 and 2007, data from internal medicine residents who entered the clinical part of the board examination in the first attempt were collected. For one candidate, scores from each examiner for each of the long case and short case were recorded. All statistical analyses were performed using statistical software SPSS version 13.0 (SPSS Inc., Chicago, USA).

Reliability study

Intraclass correlation coefficient (ICC) was used to determine the extent to which the scores given by an examiner were in agreement with one another (inter-rater reliability)[5]. The ICC of 0.6-0.8 was considered as good agreement and considered as excellent if the value was more than 0.8. Variation between the two examiners was calculated with the below formula:

$$\text{Variation (\%)} = \frac{\text{Score from the first examiner - Score from the second examiner}}{\text{Score from the first examiner + Score from the second examiner/2}}$$

Mean and standard deviation of the overall variation were presented for each of the long cases and short cases.

Validity study

The RCPT clinical examination was split into the mid-year and end-of- year examination to promote the chance for further development after an initial encounter. Correlation between mid-year and end-of-year total scores for long case and short case were assessed by Pearson method.

In 2007, the structure of examination in mid-year and end-of-year examinations were similar. If a candidate gained experience from the mid-year examination, improvement in percentage of score for he or she in the end-of-year examination was expected (construct validity). Normalized gain of each candidate was calculated with the below formula[6]:

$$\text{Normalized gain} = \frac{\text{\% End-of-year score - \% Mid-year score}}{\text{100 - \% Mid-year score}}$$

Positive normalized gain > 0.7 was considered high, 0.3-0.7 moderate, and < 0.3 as low. Negative normalized gain was considered if end-of-year score was less than the mid-year score resulting in minus value.

Mean $\pm$ standard deviation of percent variation and class gain were used to summarized the long case and short case. Corelation between mid year and end of year scores and between examiners were calculated with p-value was set at less than 0.05 for statistically significant.

Results

There were 182, 184, and 219 candidates in the years 2005-2007 respectively. The intraclass correlation (ICC) between examiners for each short case and long case in 2005 and 2006 are shown in Table 1 and for the 2007 in Table 2. Most of the ICC for a long case was in a good range except for that in 2006 that was in excellent range. All of the ICC for short cases were in excellent range except for only one short case in 2006.

Variations in percentage of score between examiners in each long and short case are shown in Table 3 and Table 4 respectively. The range of variation was between 15.3 and 27.3%.

The correlation between mid-year and end-of-year scores is shown in Table 5. Although it had statistical significance for a long case, the correlation

Table 1. Intraclass correlation between examiners in 2005 and 2006

Examination	Correlation	
	2005	2006
Mid-year		
Long case 1	0.79	0.83
Long case 2	0.71	0.81
End-of-year		
Long case 3	0.69	0.83
Short case 1	0.95	0.97
Short case 2	0.86	0.80
Short case 3	0.97	0.87
Short case 4	0.91	0.92
Short case 5	0.86	0.72

Table 2. Intraclass correlation between examiners in 2007

Examination	Correlation
Mid-year	
Long case 1	0.78
Long case 2	0.73
Short case 1	0.90
Short case 2	0.82
Short case 3	0.83
End-of-year	
Long case 3	0.75
Long case 4	0.76
Short case 4	0.89
Short case 5	0.87
Short case 6	0.82

Table 3. Percentage of examiner variation for long case (LC)

Year	% Variation (mean $\pm$ SD)			
	LC1	LC2	LC3	LC4
2005	9.0 $\pm$ 7.7	9.9 $\pm$ 7.8	10.3 $\pm$ 9.2	
2006	9.2 $\pm$ 7.8	10.0 $\pm$ 7.1	9.6 $\pm$ 7.7	
2007	11.9 $\pm$ 13.8	10.9 $\pm$ 10.1	10.6 $\pm$ 9.1	9.2 $\pm$ 7.7

was rather weak. For the short case in 2007, there was no significant correlation.

Class normalized gain was negative for long cases in 2005 and 2006 (Table 6). In 2007, both of the class normalized gain for long case and short case were also negative. When considering each candidate, nearly half of them had negative individual normalized gain. For those who had positive normalized gain, most had only low to medium gain.

Discussion

Written examination is mainly to evaluate the medical knowledge while the clinical examination aims to assess skills and attitudes. When measuring the training outcomes of internal medicine residents, reliable and valid assessment tools must be used to distinguish between those with adequate clinical competence and those without. For the RCPT clinical examination which involved multiple unstandardized long cases and short cases, each candidate encounter the patients under the observation of two independent examiners. This high stakes tests needs a verification for its reliability and validity to ensure fairness for all candidates.

High intraclass correlation with narrow variation range in the present study indicates that the scores given by different observers were very similar especially for short cases. Although the RCPT instructed all examiners to rate each candidate independently, the authors cannot rule out the influence of a senior or specialist examiner on their partner's ratings which can result in high or excellent inter-rater reliability.

The RCPT long case seems to have a reasonable degree of face validity because it allows the assessment of candidate's ability to integrate information gathered from history taking, physical examination, and laboratory interpretation for formulating the diagnosis and management plan in different clinical vignettes. But the weak correlation in 2005 and 2006 and the failure in nearly half of the

Table 4. Percentage of examiner variation for short case (SC)

Year	% Variation (mean $\pm$ SD)					
	SC1	SC2	SC3	SC4	SC5	SC6
2005	9.6 $\pm$ 9.8	15.9 $\pm$ 17.5	6.9 $\pm$ 8.4	8.3 $\pm$ 11.8	10.1 $\pm$ 13.3	
2006	8.3 $\pm$ 10.6	11.6 $\pm$ 11.6	8.0 $\pm$ 8.2	11.6 $\pm$ 10.4	11.4 $\pm$ 10.6	
2007	11.2 $\pm$ 11.1	12.7 $\pm$ 11.4	13.5 $\pm$ 13.8	9.7 $\pm$ 7.5	10.2 $\pm$ 8.4	13.1 $\pm$ 12.9

candidates to improve their scores in 2007 after being examined 6-month apart contribute to low construct validity of the presented current outcome evaluation methods. Strengths and weaknesses of the long case for assessment of clinical competence have been criticized with some suggestions for cautious use[7,8].

For the short case examination with focus evaluation points in a certain period of time, even though it had higher inter-rater reliability than long case, the authors cannot demonstrate its construct validity in 2007 where the structure of mid-year and end-of-year examination was identical. However, higher expectation of examiners for final assessment of the candidates may lead to a tough rating during the end-of-year examination. Future development of a system for examiner training and calibration is warranted for improving quality of the RCPT clinical examination[9,10].

Some certain limitations in the present study deserved mentioning. The authors have used a parallel two-way random effects model for calculating intraclass correlation which required random assignments of examiners and patients to the candidates. The RCPT did not have a systematic method of assignment, so the inter-rater reliability may be over or under estimated. The authors also did not use the equating methods to compensate the candidates who were tested by more stringent examiners (hawks) or more liberate examiners (doves). Finally, the authors do not have evidence to demonstrate the content and predictive validity in the present clinical examination. Whether the long case and short case test's contents correlate with the construct it is intended to measure, and also the relationship with another instrument or outcome should also be further verified.

Table 5. Correlation between mid-year and end-of-year scores

Year	Scores	Correlation	p-value
2005	Long case	0.27	<0.001
2006	Long case	0.19	0.009
2007	Long case	0.25	<0.001
	Short case	0.08	0.23

Conclusion

Current long case and short case in the clinical examination held by the Royal College of Physicians of Thailand had acceptable inter-rater reliability, but its construct validity in terms of improving the ability of candidates after an interval of 6 months between the first and the final encounters could not be demonstrated.

References

1. Holmboe ES, Hawkins RE. Methods for evaluating the clinical competence of residents in internal medicine: a review. Ann Intern Med 1998; 129: 42-8.
2. Hamdy H, Prasad K, Williams R, Salih FA. Reliability and validity of the direct observation clinical encounter examination (DOCEE). Med Educ 2003; 37: 205-12.
3. Wilkinson TJ, Campbell PJ, Judd SJ. Reliability of the long case. Med Educ 2008; 42: 887-93.
4. Chierakul N, Danchaivijitr S, Kontee P, Naruman S. Relationship between outcome of written and clinical parts in Internal Medicine Board Certifying Examination. Siriraj Med J 2009: 61: 194-64.
5. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med 2006; 119: 166-16.
6. Hake R. Interactive-engagement vs traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. Am J Phys 1998; 66: 64-74.
7. Norcini J. The validity of long cases. Med Educ 2001; 35: 720-1.
8. Norcini JJ. The death of the long case? BMJ 2002; 324: 408-9.
9. Wass V, Jolly B. Does observation add to the

Table 6. Normalized gain from mid-year to end-of-year scores

Year	Scores	Mean ± SD for class gain (%)	Negative gain (%)	Low gain (%)	Medium gain (%)	High gain (%)
2005	Long case	-0.7 ± 38.5	45.9	30.4	22.7	1.0
2006	Long case	-7.9 ± 61.8	47.2	28.0	21.8	3.1
2007	Long case	-8.1 ± 55.1	45.6	32.0	21.5	0.9
	Short case	-9.0 ± 61.4	48.2	23.2	24.1	4.4

validity of the long case? Med Educ 2001; 35: 729-34.
10. 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.

ความน่าเชื่อถือ และความถูกต้องของการสอบรายยาว และรายสั้นในการสอบเพื่อวุฒิบัตรสาขาอายุรศาสตร์

นิธิพัฒน์ เจียรกุล, สมหวัง ด่านชัยวิจิตร, ผกา คณฑี, ชนะ นฤมาน

วัตถุประสงค์*: เพื่อให้ได้รับวุฒิบัตรเพื่อแสดงความรู้ความชำนาญในการประกอบวิชาชีพเวชกรรมสาขาอายุรศาสตร์ ผู้มีคุณสมบัติจะต้องสอบผ่านทั้งภาคทฤษฎีและภาคปฏิบัติที่จัดขึ้นต่างปีการศึกษากัน การศึกษานี้ต้องการแสดงถึงความน่าเชื่อถือ และความถูกต้องของการสอบรายยาว และรายสั้นที่เป็นส่วนสำคัญของการสอบภาคปฏิบัติเพื่อวุฒิบัตร*

วัสดุและวิธีการ*: รวบรวมข้อมูลของแพทย์ประจำบ้านจำนวน 585 คน ที่เข้าสอบภาคปฏิบัติเพื่อวุฒิบัตรเป็นครั้งแรก ระหว่างปีการศึกษา 2548-2550 ทำการศึกษาความน่าเชื่อถือระหว่างการให้คะแนนของผู้คุมสอบ และความถูกต้องในการวัดผลของการสอบทั้งในส่วนรายยาว และรายสั้น*

ผลการศึกษา*: มีความสัมพันธ์ในระดับดีถึงดีมากของคะแนนที่ได้จากผู้คุมสอบต่างคนกัน โดยมีค่าความสัมพันธ์ระหว่าง 0.71 และ 0.97 ความแปรปรวนของคะแนนอยู่ระหว่างร้อยละ 15.3 และ 27.3 ในการสอบต่างครั้งกัน มีผลคะแนนที่ได้เพิ่มขึ้นเป็นร้อยละ -0.7 ถึง -9.0 โดยผู้เข้าสอบร้อยละ 45.6 ถึง 48.2 มีผลการสอบครั้งที่สองที่ได้คะแนนลดลง*

สรุป*: ในการสอบรายยาว และรายสั้นของการสอบภาคปฏิบัติเพื่อวุฒิบัตรสาขาอายุรศาสตร์ การให้คะแนนระหว่างผู้คุมสอบ มีความสอดคล้องกันอย่างน่าเชื่อถือ แต่ไม่สามารถแสดงให้เห็นความถูกต้องของการประเมินความรู้ความสามารถทางคลินิกได้*