Hong Kong J Radiol. 2017;20:232-6 | DOI: 10.12809/hkjr1716804

ORIGINAL ARTICLE

Inter-rater Reliability of Examiners in the Hong Kong College of Radiologists' Palliative Medicine Oral Examination

R Chow[1], L Zhang[1], IS Soong[2], OWK Mang[3], LCY Lui[4], KH Wong[3], SWK Siu[5], SH Lo[6], KK Yuen[5], YSH Yau[7], KY Wong[4], C Leung[3], SY Wong[6], R Ngan[3], E Chow[1], R Yeung[2]

[1]Sunnybrook Health Sciences Centre, University of Toronto, Toronto, ON, Canada; [2]Department of Clinical Oncology, Pamela Youde Nethersole Eastern Hospital, Chai Wan, Hong Kong; [3]Department of Clinical Oncology, Queen Elizabeth Hospital, Jordan, Hong Kong; [4]Department of Clinical Oncology, Princess Margaret Hospital, Lai Chi Kok, Hong Kong; [5]Department of Clinical Oncology, Queen Mary Hospital, Pokfulam, Hong Kong; [6]Department of Clinical Oncology, Tuen Mun Hospital, Tuen Mun, Hong Kong; [7]Department of Clinical Oncology, Prince of Wales Hospital, Shatin, Hong Kong

ABSTRACT

Objective: *To analyse the inter-rater reliability of scores in the Palliative Medicine Oral Examination among examiners, among observers, and between examiners and observers.*

Methods: *The Palliative Medicine Subspecialty Board aims to train oncology specialists for palliative medicine through a 4-year accreditation programme. At the end of the programme, trainees undergo a Board Examination involving subjective ratings by examiners. Each candidate rotated through two panels during the 1-day examination; one panel involved the written dissertation and questions pertaining to symptom management (viva 1) and the other about psychosocial issue (viva 2) and ethics (viva 3). A total of 10 candidates were evaluated on the four occasions using a 10-point scale by six examiners and four observers, along with one external examiner. Intraclass correlation coefficient (ICC) was calculated to determine inter-rater reliability (concordance) among examiners, among observers, and between examiners and observers. ICC values are classified as poor (≤0.20), fair (0.21-0.40), moderate (0.41-0.60), good (0.61-0.80), and very good (0.81-1.00).*

Results: *Among examiners, concordance was overall good at different stations. Among observers, concordance was fair to very good across different stations. Between examiners and observers, concordance was fair to moderate at two stations. Across all stations, concordance was good between examiners and observers.*

Conclusion: *The inter-rater reliability was good at the Board Examination administered by the Palliative Medicine Subspecialty Board of the Hong Kong College of Radiologists. The examination is reliable in accrediting practitioners for subspecialty certification.*

Key Words: *Oncologists; Palliative medicine*

Correspondence: *Dr R Yeung, Department of Clinical Oncology, Pamela Youde Nethersole Eastern Hospital, 3 Lok Man Road, Chai Wan, Hong Kong.*
Email: yeungmwr@ha.org.hk

Submitted: 31 Aug 2016; Accepted: 2 Feb 2017.

Disclosure of Conflicts of Interest: All authors have disclosed no conflicts of interest.

中文摘要

考官在香港放射科醫學院紓緩醫學口試評分者間的可靠性

R Chow、張麗瑩、宋崧、孟偉剛、呂卓如、黃錦洪、蕭偉君、魯勝雄、袁國強、邱秀嫻、黃家仁、梁偉濂、王韶如、顏繼昌、E Chow、楊美雲

目的：分析考官之間、觀察員之間以及考官與觀察員之間的紓緩醫學口試評分者間的可靠性。

方法：紓緩醫學專業委員會旨在通過一個4年認證計劃為腫瘤專科醫師進行紓緩醫學培訓。在課程結束時，學員需要通過考官的主觀評級考試。每個考生在一天的考試中輪流接受兩組考官的評核：一組涉及有關書面論文及症狀處理（viva 1），另一組涉及心理社交（viva 2）和倫理（viva 3）。共10名考生被六名考官和四名觀察員以及一名外部考官評估。使用內部相關系數（ICC）評估考官之間、觀察員之間以及考官和觀察員之間的評分者間可靠性（一致性）。ICC值分為差（≤0.20）、一般（0.21-0.40）、中等（0.41-0.60）、好（0.61-0.80）和非常好（0.81-1.00）。

結果：考官之間的一致性在不同站總體為好。觀察員之間的一致性在不同站為一般至非常好。考官和觀察員之間的一致性在兩個站為一般至中等。所有站之間，考官和觀察員間的一致性為好。

結論：香港放射科醫學院紓緩醫學專科委員會的紓緩醫學口試評分者間的可靠性好。該口試能可靠認證醫師的專業資格。

INTRODUCTION

The Palliative Medicine Subspecialty Board, established in 2002, is a subsidiary of the Hong Kong College of Radiologists. Its mission is to provide palliative medicine training for clinical oncologists to deliver the best comprehensive care. Its training programme consists of a 4-year accreditation programme: 2 years of Higher Specialist Training in Clinical Oncology and another 2 years of Palliative Medicine Subspecialty Training.[1]

At the end of the programme, trainees are required to sit for the Board Examination, which consists of a dissertation appraisal examination and an oral examination. The dissertation varies from 5000 to 20,000 words, and the 1-day oral examination focuses on dissertation content in addition to various other aspects of palliative medicine. In March 2016, the fourth Palliative Medicine Board oral examination took place. The candidates went through four stations and were subjectively evaluated by examiners and observers.

Inter-rater reliability has shown to be a topic of clinical research in recent years. The inter-rater reliability comparing checklist and global scoring for objective structured clinical examinations has been reported.[2] While others have done similar research,[3,4] the subjective rating invites research for inter-rater reliability.[5,6] This study aimed analyse the inter-rater reliability of scores in the Palliative Medicine Oral Examination among examiners, among observers, and between examiners and observers.

METHODS

In the fourth Board Examination, a total of 10 candidates were examined in two panels during the 1-day examination by six examiners (examiners 2-7) and four observers (observers A-D), along with one external examiner from Canada. Panel 1 contained questions on the written dissertation and viva 1 pertaining to symptom management, whereas panel 2 consisted of psychosocial issue (viva 2) and ethics (viva 3). The examiners suggested potential questions for viva 1-3 to the pool. They met with the external examiner and the observers the day before the oral examination to discuss all the questions from the pool and then select the final three questions for the symptom management, psychosocial issue, and ethics. They also discussed what the acceptable answers would be. The decision of the final examination questions and discussion was kept in strict confidence.

Each panel had two local examiners and two observers. To be fair to all candidates, extra caution was exercised

to make sure the examiner and the candidate from each panel did not come from the same centre. The external examiner observed each examiner and each candidate at least once.

Each assessor (two examiners and two observers for each panel) rated the performance of candidates on a 10-point scale, with one being the lowest score. A pass was defined as a score of 5 or above. In panel 1, assessors gave two scores: dissertation and clinical oncology (viva 1). In panel 2, assessors completed two evaluations pertaining to ethics and role-play (viva 2 and 3).

Intraclass correlation coefficient (ICC)[7] was used to estimate the concordance or inter-rater reliability among examiners, among observers, and between examiners and observers. ICC ranges from 0 to 1, with 1 indicating perfect concordance. The interpretation criteria for ICC values are poor (≤0.20), fair (0.21-0.40), moderate (0.41-0.60), good (0.61-0.80), and very good (0.81-1.00).[8]

RESULTS

Among examiners, concordance at dissertation, viva 1, viva 2, and viva 3 were 0.514 (moderate), 0.220 (fair), 0.736 (good), and 0.685 (good), respectively, with an overall ICC of 0.637 (good). Examiners at the dissertation station had individual ICC ranging from 0.250 (fair) to 0.955 (very good). Viva 1 ranged from 0.500 (moderate) to 0.716 (good), viva 2 from 0.538 (moderate) to 0.743 (good), and viva 3 from 0.483 (moderate) to 0.703 (good) [Table 1].

Among observers, concordance was fair for dissertation (ICC = 0.380) and viva 1 (ICC = 0.374), moderate for viva 2 (ICC = 0.530) and very good for viva 3 (ICC = 0.838). Overall, across all stations, inter-rater reliability was fair (ICC = 0.318). There was, at the dissertation station, poor concordance between the two observers with respect to candidates' performance (ICC = 0.077). Otherwise, individual ICC values ranged from 0.286 (fair) to 0.819 (very good) [Table 2].

Between examiners and observers, moderate concordance was noted at the dissertation station (ICC = 0.567). Viva 1 had fair concordance (ICC = 0.349), while viva 2 and viva 3 had good concordance (ICC = 0.736 and 0.778, respectively). Across all stations, there was good concordance between examiners and observers (ICC = 0.619) [Table 3].

All 10 candidates passed the 1-day oral examination. Inter-rater reliability in the pass-fail outcome of the Board Examination, by station, was very good—96%. In all but two stations, examiners agreed to pass candidates; the two exceptions occurred where the two examiners scored the candidate with four and six, but the combined score was 10 of 20 with the pass. Overall, the external examiner agreed with all assessments of local examiners to pass the candidates, further confirming inter-rater agreement.

DISCUSSION

This is the first study to examine inter-rater reliability of examiners in a medical board examination setting in Hong Kong. Universally, board examinations usually

Table 1. Inter-rater reliability among examiners.

Station	Intraclass correlation coefficient (95% confidence interval)	Interpretation	Range of difference in scores
Dissertation	0.514 (0-0.866)	Moderate	
Examiners 4 & 5	0.955 (0.798-0.997)	Very good	0-0.5
Examiners 1 & 2	0.341 (0-0.942)	Fair	0-2
Examiners 3 & 6	0.250 (0-0.981)	Fair	1
Viva 1	0.220 (0-0.791)	Fair	
Examiners 4 & 5	0.716 (0-0.982)	Good	0.5-1
Examiners 1 & 2	0.688 (0-0.888)	Good	1-2
Examiners 3 & 6	0.500 (0-0.978)	Moderate	1-2
Viva 2	0.736 (0.446-0.920)	Good	
Examiners 1 & 3	0.538 (0-0.968)	Moderate	0-2
Examiners 4 & 6	0.743 (0.252-0.966)	Good	1-2
Viva 3	0.685 (0.333-0.906)	Good	
Examiners 1 & 3	0.483 (0-0.966)	Moderate	1-2
Examiners 4 & 6	0.703 (0.191-0.960)	Good	1-2
Overall	0.637 (0.222-0.892)	Good	

Table 2. Inter-rater reliability among observers.

Station	Intraclass correlation coefficient (95% confidence interval)	Interpretation	Range of difference in scores
Dissertation	0.380 (0-0.818)	Fair	
Observers C & D	0.077 (0-0.818)	Poor	0-1
Observers A & B	0.547 (0-0.934)	Moderate	0-2
Viva 1	0.374 (0-0.831)	Fair	
Observers C & D	0.417 (0-0.929)	Moderate	0-2
Observers A & B	0.308 (0-0.924)	Fair	0-2
Viva 2	0.530 (0.078-0.0861)	Moderate	
Observers C & D	0.286 (0-0.905)	Fair	0-1
Observers A & B	0.579 (0-0.940)	Moderate	0-2
Viva 3	0.838 (0.658-0.955)	Very good	
Observers C & D	0.778 (0.301-0.973)	Good	0-1
Observers A & B	0.819 (0.398-0.977)	Very good	0-1
Overall	0.318 (0.022-0.707)	Fair	

Table 3. Inter-rater reliability between examiners and observers.

Station	Intraclass correlation coefficient (95% confidence interval)	Interpretation
Dissertation	0.567 (0-0.876)	Moderate
Viva 1	0.349 (0-0.813)	Fair
Viva 2	0.736 (0.378-0.924)	Good
Viva 3	0.778 (0.501-0.935)	Good
Overall	0.619 (0.127-0.889)	Good

disclose results as 'pass' or 'fail' to candidates, as the aim is to accredit candidates who have displayed the minimum requirements to be a safe practitioner. The 96% agreement in pass-fail results confirms the reliable nature of the examination.

Among the examiners, inter-rater reliability was good at panel 2, for both viva 2 and viva 3. However, it was moderate in the dissertation station and fair in panel 1 viva 1. It is important to note however that moderate concordance was a result of 1- or 2-point differences out of 10 between the examiners; good to very good concordance was only statistically identified when examiners gave identical or almost-identical (0.5-point difference) scores. The same also applies in measuring the inter-rater reliability among observers and between examiners and observers. The variability at the dissertation station may be accounted for by the nature of the assessment; the dissertation comprised a PowerPoint presentation and question-and-answer period. Although sub-par concordance was also observed in the viva 1 station, it is important to note that individual ICC was moderate or good; the fair concordance was a result of the overall ICC calculated across several raters with a greater range of variability.

Among the observers, the inter-rater reliability was fair in both dissertation and panel 1 viva 1, moderate in panel 2 viva 2, and very good in panel 2 viva 3. Greater variability was observed across the observers (when compared to the examiners) and this may be likely due to less experience. This is acceptable in this setting as the aim is for observers to learn the assessment format and hopefully achieve good concordance when they become examiners in the subsequent examination. It will be of interest to follow these observers on their inter-rater reliability when they become examiners in future examinations. It is encouraging to see that observers and examiners had good concordance across all stations. Overall, the 10-point scale has proven to be a reliable tool in this examination setting. Previous study also showed high inter-rater reliability of the scores provided by examiners. The inter-rater reliability was higher for global ratings than for checklist scores.[2]

This study was not without limitations. The small number of candidates led to lower power in the statistical analysis. However, this was unavoidable. There are usually 10 or less oncologists participating in the palliative medicine accreditation programme. However, because of our decreased sample sizes after breakdown, the 95% confidence interval of ICC in each level of breakdown has much wider ranges, comparing to all examiners (0.637 with 0.222-0.892). Although we have very good / moderate ICC in dissertation / other panels, it is more reliable to consider good ICC when we use all examiners, due to sample size issue.

CONCLUSION

This study highlights the strong inter-rater reliability

between the examiners at the Board Examination administered by the Palliative Medicine Subspecialty Board of the Hong Kong College of Radiologists. The pre-examination meeting among the examiners and observers has definitely helped to improve the inter-rater reliability. The examination is reliable in accrediting practitioners for the subspecialty certification.

REFERENCES

1. Hong Kong College of Radiologists. Subspecialty Training: Palliative Medicine; 2011. Available from: https://www.hkcr.org/templates/OS03C00336/pdf/TrainingGuidelines_PallaitiveMedicine-withAppendix_Version201111.pdf. Accessed 25 July 2016.
2. Malau-Aduli BS, Mulcahy S, Warnecke E, Otahal P, Teague PA, Turner R, et al. Inter-rater reliability: comparison of checklist and global scoring for OSCEs. Creat Educ. 2012;3:937-42. cross ref
3. Cohen DS, Colliver JA, Robbs RS, Swartz MH. A large-scale study of the reliabilities of checklist scores and ratings of interpersonal and communication skills evaluated on a standardised-patient examination. Adv Health Sci Educ Theory Pract. 1996;1:209-13. cross ref
4. Cunnington JP, Neville AJ, Norman GR. The risks of thoroughness: reliability and validity of global ratings and checklists in an OSCE. Adv Health Sci Educ Theory Pract. 1996;1:227-33. cross ref
5. Zimmermann C, Burman D, Bandukwala S, Seccareccia D, Kaya E, Bryson J, et al. Nurse and physician inter-rater agreement of three performance status measures in palliative care outpatients. Support Care Cancer. 2010;18:609-16. cross ref
6. Chow R, Chiu N, Bruera E, Krishnan M, Chiu L, Lam H, et al. Inter-rater reliability in performance status assessment among health care professionals: a systematic review. Ann Palliat Med. 2016;5:83-92. cross ref
7. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86:420-8. cross ref
8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-74. cross ref