

**DISTANCE RATER-TRAINING PROGRAM DO IMPROVE INTER-RATER
RELIABILITY
IN MEDICAL EDUCATION**

Stanley SL. Tsai

National Taiwan University Hospital
bientsai@gmail.com

Abstract

Background: Many well-constructed and validated evaluation tools have been used for clinical performance assessment (ex. mini-clinical evaluation exercise) in medical education for years. Rater training for clinical performance assessment has long been discussed, and the best way to achieve consensus remains controversial. Objectives: To investigate the effect of distance rater- training program of which applied distance learning technology. Method: 24 intern doctors who had already undergone half year of internship completed an 8-minute interview of a simulated case with abdominal pain portrayed by well-trained standardized patients in May, 2012. Intern doctors' performance was captured on videotape. 17 clinical faculties were invited to rate the intern doctors and 12 of 17 voluntarily participated in the distance rater-training program. One week after distance rater-training program, 17 faculties were assigned to two groups to rate intern doctors performance: Group A- 5 trained raters and 5 un-trained raters; Group B- 7 trained raters. The scores of intern doctors were calculated and inter-rater reliability was examined with Pearson analysis. Levene test was performed on the difference between two groups. Distance Rater-training Program: An internet website for rater-training was established. Training context included a 40-minute lecture, videos with checklist for rating practice, and an internet discussion forum. Results: The inter-rater reliability of Group A is 0.623 ($Z=0.803$) and 0.755 ($Z=1.298$) of Group B. The difference of Z scores of two groups was examined by Levene test which showed a significant difference ($p=0.002$). Conclusions: The distance

rater-training program has demonstrated a way to provide a better inter-rater reliability of a trained rater group.

Keywords: clinical performance assessment, inter-rater reliability.

Instead of memorizing tons of medical knowledge, the goals of medical education are to cultivate physicians capable of applying their medical skills to meet the needs of patients and society, and to create a high quality medical care environment. (ABIM, 2001; ACGME, 2003) In the past decade, clinical performance assessment therefore drew more attention in medical education than ever. (Norcini, 2003; Holmboe, 2004) Research found scores from the American Board of Internal Medicine “long case” clinical evaluation exercise (CEX) to be unreliable due to large inter-case and inter-rater variance. (Kroboth, 1992) Rater training for clinical performance assessment became a challenging issue as a result, and the best way to achieve consensus remains controversial. (Wilkinson, 2003; Beckman, 2004; Downing, 2005; Cook, 2009)

However, tremendous advances in technology have made many impossible things practical. Inter rater-reliability, usually the most challenging part of the clinical performance assessment, could be improved by innovative technology or not is already becoming a promising issue for medical educators.

1. OBJECTIVE

This study aims to investigate the effect of distance rater-training program of which applied distance learning technology.

2. METHOD

8 clinical educators were invited from different departments of National Taiwan University Hospital and were asked to develop an 8-minute simulated case of abdominal pain which is the most common and basic symptom an intern doctor has to be capable of dealing with. 15 items of the checklist was designed to evaluate intern doctor interviewing skill. Case development was completed in January, 2012 and was pilot tested by examining 3 first-year resident doctors interviewing a well-experienced

standardized patient(SP). These 3 resident doctors were intern doctors 6 months before pilot-testing. Pilot-testing was video-taped and reviewed by the 8 educators, and a rating reference was created. Minor revision was made to make simulated case more practical and validated.

17 NTUH clinical faculties were invited to rate the intern doctors. 12 of 17 faculties voluntarily participated in the distance rater-training program. Training program consisted of an internet website which included a 40-minute lecture introducing the concept of clinical performance assessment and factors affecting inter-rater reliability. Pilot-testing videos and the rating checklist were all uploaded to the website. 12 participants were provided individual identity and password and they have to log in to practice rating at their convenience within one week. After practicing rating, an internet discussion forum was available for participants and the forum tagged with all participants' test-scoring results. By reviewing the different results from participants, discussion was raised on the forum to establish the consensus on performance standard of each item of checklist.

One week after distance rater-training program, 17 faculties were assigned to two groups to rate intern doctors performance: Group A- 5 trained raters and 5 un-trained raters; Group B- 7 trained raters. 24 NTUH intern doctors who had already undergone 10 months of internship completed an 8-minute interview of that simulated case portrayed by well-trained standardized patients in May, 2012. The scores of intern doctors were calculated and inter-rater reliability was examined with Pearson analysis. Levene test was performed on the difference between two groups.

3. RESULTS

The inter-rater reliability of Group A is 0.623 ($Z=0.803$) and 0.755 ($Z=1.298$) of Group B. The difference of Z scores of two groups was examined by Levene test which showed a significant difference ($p=0.002$).

4. CONCLUSIONS

The distance rater-training program provides a convenient way for establishing performance rating consensus and also demonstrates a better way to improve inter-rater reliability.

REFERENCES

ABIM, American Board of Internal Medicine (2001). Portfolio for Internal Medicine Residency Programs. Philadelphia: American Board of Internal Medicine.

ACGME, Accreditation Council for Graduate Medical Education (2003). Outcome Project: The General Competencies. USA: Accreditation Council for Graduate Medical Education.

Holmboe E, S., Hawkins R, E., & Huot S, J. (2004). Effects of training in direct observation of medical residents' clinical competence: a randomized trial. *Annals of Internal Medicine*, 140(11), 874–881.

Beckman T, J., Ghosh A, K., Cook D, A., Erwin P, J., & Mandrekar J, N. (2004). How reliable are assessments of clinical teaching? *Journal of General Internal Medicine*, 19(9), 971–977.

Cook D, A., Duprase D, M., Neckman T, J., Tomas K, G. (2009). Effect of rater training on reliability and accuracy of mini-CEX scores: A randomized, controlled trial. *Journal of General Internal Medicine*, 24(1), 74-79.

Downing S, M. (2005). Threats to the validity of clinical teaching assessments: What about rater error? *medical education*, 39(4), 353-355.

Kroboth F, J., Hanusa B, H., & Parker S. (1992). The inter-rater reliability and internal consistency of a clinical evaluation exercise. *Journal of General Internal Medicine*

Medicine, 7(2), 174–179.

Norcini J, J., Blank L, L., Duffy F, D., & Fortna G, S. (2003). The mini-CEX: a method for assessing clinical skills. *Annals of Internal Medicine*, 138(6), 476–481.

Wilkinson T, J., Frampton C, M., & Thompson-Fawcett M. (2003). Objectivity in objective structured clinical examinations: checklists are no substitute for examiner commitment. *Academic Medicine*, 78(2), 219-223.