Tzu Chi Medical Journal 24 (2012) 155–156

Contents lists available at SciVerse ScienceDirect

Tzu Chi Medical Journal

journal homepage: www.tzuchimedjnl.com

Medical Education

Effectiveness of rater consensus training in objective structured clinical examination using Kolb's experiential learning

Shao-Yin Chu [a,b], Yu-Ling Lan [c], Sheng-Po Kao [a,b], Tsung-Ying Chen [a,b], Ming-Chen Hsieh [a,b,*]

[a] Department of Medical Education, Buddhist Tzu Chi General Hospital, Hualien, Taiwan
[b] Department of Medicine, College of Medicine, Tzu Chi University, Hualien, Taiwan
[c] Department of Counseling and Clinical Psychology, National Dong Hwa University, Hualien, Taiwan

ARTICLE INFO

Article history:
Received 28 September 2011
Received in revised form
3 October 2011
Accepted 9 January 2012

A successful, valid, reliable, and defensible high-stakes objective structured clinical examination (OSCE) depends on many factors, such as the test contents, the quality of the performance of standardized patients (SPs), and the consensus of judgments by raters on student OSCE performance. Higher inter-rater reliability of clinician examiners is demonstrated when assessments are recorded on structured forms and examiners participate in station construction [1]. The consistency among raters' judgments was affected by the level of rater training [2,3]. The literature shows few rater training methods or models [4]. This article describes the effectiveness of rater consensus training in high-stakes OSCE using Kolb's cycle of experiential learning [5]. According to Kolb [6], learning is the process whereby knowledge is created through the transformation of experience.

Population: Raters were recruited from six hospitals, and 60% of them had rater experience in clinical performance examinations.

Design: A half-day rater consensus-training workshop was carried out 1 day before administrating a high-stakes OSCE. An immediate response system (IRS) with 15 major questions was designed to test raters' concrete experience and their intent to apply that knowledge learned from e-learning. Two rater trainers bidirectionally discussed the immediate test results. To foster reflective observation and to decrease raters' scoring variations, two case-specific raters were grouped for each OSCE station, and a modified Delphi method was used to promote the greatest consensus in the structured rating scale [7]. Participants engaged in two rating cycles with full discussion to reach an agreement. Table 1 shows an overview of the course. Subsequently, we assessed participants' learned knowledge and skills during a 2-day high-stakes OSCE with eight stations with SPs.

Data analysis: We assessed raters' satisfaction with the workshop and self- assessment of their own rating qualities on the high-stakes OSCE with questionnaires after they completed the examination. We examined rater consensus in two ways, i.e., by inter-rater reliability and by scoring differences between two raters on each OSCE station. We determined the inter-rater reliability by the Pearson product-moment correlation coefficient. We determined the rater scoring differences by the related-samples *t*-test using a 0.05 level of significance.

A total of 49 raters were recruited to participate in this high-stakes OSCE. All of them completed an e-learning course before attending the OSCE rater-training workshop. We used IRS assessment to evaluate whether these raters had learned effective knowledge from the e-learning course. Immediate discussion of assessment results and case-specific rater consensus training guided reflective observation, discussion, and reconceptualization. Self-assessed rating qualities on the high-stakes OSCE received a mean score of 4.4 points on a five-point Likert scale. Up to 90% of raters reported a willingness to attend the next rater's performance. Correlations between raters' judgments on the eight high-stakes OSCE stations ranged from 0.34 to 0.69, which indicated moderate to high inter-rater reliability across all OSCE stations. We found no significant score differences between two raters' judgments on the same OSCE student performance on six out of eight stations. These results indicate that raters did reach at least moderate agreement when evaluating students' performances on most high stakes OSCE stations (Table 2).

E-learning is an effective method to train raters. The IRS supports rapid determination of raters' learned knowledge and

* Corresponding author. Department of Medical Education, Buddhist Tzu Chi General Hospital, 707, Section 3, Chung-Yang Road, Hualien, Taiwan. Tel.: +886 3 8561825; fax: +886 3 8563532.
E-mail address: jeany.jeany@msa.hinet.net (M.-C. Hsieh).

doi:10.1016/j.tcmj.2012.02.003

Table 1
Using Kolb's experiential learning cycle in rater training.

Component of Kolb's cycle of experiential learning	Rater training process in high-stake OSCE
Concrete experience	1. e-learning in advance adds to knowledge about high-stake OSCE 2. Assessment of what is known
Reflective observation and abstract conceptualization	1. Discuss immediately the concrete experience using the immediate response system 2. Discuss rating scores for two cycles under modified Delphi consensus training 3. Observation of inconsistent rater behaviors
Active experimentation	1. Rating during consensus training 2. Rating at high-stake OSCE

OSCE = objective structured clinical examination.

Table 2
Inter-rater reliability among each station.

Station	Paired *t*-test				Pearson correlation	
	Mean ($n = 58$)	CI	SD (95%)	*p* value *	*r*	*p* value **
1. Leg pain	−0.17	−1.11 ~ 0.76	3.560	0.71	0.55	0.000
2. Headache	0.17	−1.06 ~ 1.40	4.676	0.78	0.49	0.000
3. Dizziness	0.97	−2.56 ~ 0.02	3.770	0.06	0.56	0.000
4. Palpitation	0.43	−0.33 ~ 1.19	2.878	0.26	0.70	0.000
5. Pancreatitis	−0.62	−1.86 ~ 0.62	4.720	0.32	0.45	0.000
6. Menopause	0.50	−0.38 ~ 1.38	3.336	0.26	0.70	0.000
7. Cough	0.28	−0.53 ~ 1.08	3.048	0.49	0.65	0.000
8. H1N1	−0.88	−1.73 ~ −0.02	3.250	0.04	0.34	0.009

*value obtained from the paired *t*-test ($\alpha = 0.05$).
**value obtained from the Pearson correlation ($\alpha = 0.01$).
CI = confidence interval; SD = standard deviation.

concrete experience, and trainers can immediately fill in gaps in knowledge. We found behavior level changes, as determined by the self-evaluation, to be high. However, raters' scoring variations remained high on two stations in this OSCE examination. Further investigation is needed to determine the reasons contributing to these raters' persistent scoring variations. Nonetheless, these early results support the idea that Kolb's cycle of experiential learning model can facilitate rater consensus training in a high-stakes OSCE setting.

Acknowledgments

We highly appreciate the advice of Dr. David Hirsh from the Department of Medicine, Cambridge Health Alliance, Harvard Medical School, Boston, MA, USA, on this article's publication, and Gui-Ling Kao of the Department of Medical Education, Buddhist Tzu Chi General Hospital, for her contributions in data collection and analysis.

References

[1] Valentino J, Donnelly MB, Sloan DA, Schwartz RW, Haydon III RC. The reliability of six faculty members in identifying important OSCE items. Acad Med 1998; 732:204–5.
[2] Martin JA, Reznick RK, Rothman A, Tamblyn RM, Regehr G. Who should rate candidates in an objective structured clinical examination? Acad Med 1996; 712:170–5.
[3] van der Vleuten CP, van Luyk SJ, van Ballegooijen AM, Swanson DB. Training and experience of examiners. Med Educ 1989;233:290–6.
[4] Liao SC, Hunt EA, Chen W. Comparison between inter-rater reliability and inter-rater agreement in performance assessment. Ann Acad Med Singapore 2010; 398:613–8.
[5] DeCoux VM. Kolb's learning style inventory: a review of its applications in nursing research. J Nurs Educ 1990;295:202–7.
[6] Kolb D. Experiential learning: experience as the source of learning and development. Englewood Cliffs, New Jersey: Prentice Hall; 1984.
[7] Cialkowska M, Adamowski T, Piotrowski P, Kiejna A. What is the Delphi method? Strengths and shortcomings. Psychiatr Pol 2008;421:5–15.