EDUCATION BRIEF

A Pilot Comparison of In-Room and Video Ratings of Team Behaviors of Students in Interprofessional Teams

Désirée Lie, MD, MSED, Regina Richter-Lagha, PhD, Sae Byul (Sarah) Ma, PharmD
Keck School of Medicine, University of Southern California, Los Angeles, California
Submitted April 25, 2017; accepted December 20, 2017; published June 2018.

Objective. To examine concordance between in-room and video faculty ratings of interprofessional behaviors in a standardized team objective structured clinical encounter (TOSCE).

Methods. In-room and video-rated student performance scores in an interprofessional 2-station TOSCE were compared using a validated 3-point scale assessing six team competencies. Scores for each student were derived from two in-room faculty members and one faculty member who viewed video recordings of the same team encounter from equivalent visual vantage points. All faculty members received the same rigorous rater training. Paired sample t-tests were used to compare individual student scores. McNemar’s test was used to compare student pass/fail rates to determine the impact of rating modality on performance scores.

Results. In-room and video student scores were captured for 12 novice teams (47 students) with each team consisting of students from four professions (medicine, pharmacy, physician assistant, nursing). Video ratings were consistently lower for all competencies and significantly lower for competencies of roles and responsibilities, and conflict management. Using a criterion of an average score of 2 out of 3 for at least one station for passing, 56% of students passed when rated in-room compared with 20% when rated by video.

Conclusion. In-room and video ratings are not equal. Educators should consider scoring discrepancies based on modality when assessing team behaviors.

Keywords: interprofessional education, team objective structured clinical encounter, assessment, rating modality, equivalence

INTRODUCTION

Assessment plays a vital role in competency-based health professions education. Simulation-based assessments are increasingly used in pharmacy education. The Objective Structured Clinical Examination (OSCE) with trained standardized patients (SPs) simulating actual patients has become a standard in evaluating clinical skills. The OSCE is defined as “an approach to the assessment of clinical competence in which the components of competence are assessed in a planned or structured way.” The OSCE is now used in as a component of some high stakes and licensure examinations, including the Canadian Pharmacist Qualifying Examination and the United States Medical Licensing Examination. With interprofessional education (IPE) becoming an accreditation standard for most health professions, including pharmacy, the same approach is now available for the assessment of team behaviors, using Team Objective Structured Clinical Encounters (TOSCEs).

In an OSCE station, usually lasting 5-15 minutes, a student is assessed by an in-room trained rater (a faculty member or a SP), who completes competency-based checklists or scales (live rating). Rating with video recordings (video ratings) is an alternative when in-room live rating is not feasible. The TOSCE assesses teamwork competencies similarly to an OSCE. But unlike a traditional OSCE, faculty raters have to simultaneously observe and rate multiple students interacting with one another and with the SP in one encounter, complicating the task of rating. TOSCE stations are typically longer than OSCE stations, taking 25 to 30 minutes. Like OSCEs, there is an incentive to use video recordings to rate students participating in TOSCEs because of limited resources to conduct in-room observations. Little is known about the equivalency of student performance scores using video-based ratings compared with in-room ratings. There is, therefore, a need to address modality of rating as a potential source of bias in TOSCE performances.

Corresponding Author: Sae Byul (Sarah) Ma, Keck School of Medicine, University of Southern California, 1000 S. Fremont Ave., Unit 7, Building A-11, Room 11166, Alhambra, CA 91803-8897. Tel: 626-457-3465. E-mail: saebyul.ma@med.usc.edu
METHODS

The study was conducted at the University of Southern California in Los Angeles and involved students from four health professions (pharmacy, physician assistant, medicine and nursing).

Eligible students were from the preclinical or clinical phases of training. Students were informed that the TOSCE was a formative interprofessional assessment, ratings would be de-identified, and no results would be shared with faculty or administrators. For the in-room rating, 16 volunteer faculty members were recruited from the same four professions via an email listserv of an IPE committee. The criterion for participation was previous experience evaluating students in clinical settings. The video ratings were completed by an experienced clinician rater and trainer with 20 years of educational evaluation and research experience.

A two-station TOSCE was designed with each team seeing two SPs in succession. Each student would have two sets of individual ratings, one for each station. A pair of in-room faculty raters was assigned to each team. The raters sat 8 feet away from the team, facing all four students who were seated in a half circle facing the SP across a small table. The SP’s face was partially visible to the raters. Raters and students were instructed not to move from their seats. Video recordings were captured with the camera positioned between the two raters. The in-room and video raters had a similar visual perspective.

Students were assigned to new teams just before the TOSCE. For each station, the student team was instructed to assess the SP and prepare the case for presentation to an attending provider. The two stations (one involving a patient with diabetes, the other with chronic pulmonary disease) were at the same level of difficulty per the clinical faculty who wrote the cases. Each station lasted 25 minutes: 5 minutes for a pre-huddle, 15 minutes with the SP and 5 minutes for a post-huddle. Raters were present for all 25 minutes, and were given 5 minutes between stations to complete their rating forms.

Lie and et al demonstrated that in-room faculty raters could accurately and reliably score four students simultaneously in a 25-minute encounter. The 16 in-room faculty raters received an email link to a training video and the rating scales one week prior to the event. They received one hour of in-person training as a group, prior to being assigned to their TOSCE student team. The video faculty rater received the same rater training and had prior rating experience as an in-room faculty rater demonstrating high inter-rater reliability as compared with other raters. To closely simulate conditions of in-room rating, the video rater viewed each team encounter once and did not replay any video when scoring students.

Rating Scale. The McMaster-Ottawa scale addresses six interprofessional competencies: communication, collaboration, roles and responsibilities, patient-centered approach, conflict management, and teamwork, with an additional global score (Table 1). The scale’s internal consistency for scoring ranges from 0.73 to 0.87. The scale was modified from 9 points to 3 points with descriptive behavioral anchors without compromising its psychometric properties. The modification with competency-based scores of 1 (below expected), 2 (at expected) and 3 (above expected), allowed for consistent and replicable rater training and scoring. The scale was applied to scoring individual student (reliability coefficient = .75) but not team performance because of low reliability (.55) for scoring team performance. While the scale has largely been used formatively to provide feedback, rating accuracy was also investigated by comparing pass/fail rates between in-room faculty and the video rater. To do so, a passing score was defined as achieving an average score of 2 (at expected) across all six competencies (excluding the global score) for at least one of two TOSCE stations.

Ratings for students were completed independently on paper by each in-room faculty. Video ratings were also completed on paper. Ratings for each student and team were then entered into Microsoft Excel (Microsoft, Redmond, WA) and analyzed using SPSS version 23 (IBM Corp., Armonk, NY). To determine potential differences based on rating modality (in-room vs. video), the correlation of individual student scores for scale items between the average of the two in-room raters and the video rater were examined. A paired sample t-test was performed to determine potential differences in ratings of individual students between modality (in-room or via video). McNemar’s test was conducted to compare pass/fail rates by modality.

Both in-room and video raters were trained similarly to ensure standardization. Therefore, the degree of training for applying the scale was considered to be equivalent. For the purpose of this study, variation in scores attributable to rater is assumed to be a result of differences in modality of rating, and not rater experience.

RESULTS

Sixty-three students in 16 teams participated in the TOSCE. Sixteen faculty members from the four professions
and the video rater received the same rater training. All 16 in-room faculty members submitted their independent ratings. There were non-significant differences in scoring of each student between the two in-room raters at each of the two stations, confirming high inter-rater reliability. Video recordings were successfully made for 12 teams (47 students), but videos of four teams (16 students) were not captured due to technical problems. One faculty who was not present at the TOSCE, rated the videos of students from the 12 teams for both SP stations. There were no differences in student scores by age, gender, profession, or training (pre-clinical vs. clinical). There was a statistically significant score difference between students who reported having a prior interprofessional experience compared with those who reported having no experience in station 1, \( p = .007 \), and station 2 \( p = .029 \). Although score differences between professions were not significant, nursing students, who more frequently reported having no prior interprofessional experiences, overall scored the lowest on the TOSCE. On average, video ratings produced lower student scores for all scale items (Table 1). Scale items with the largest difference in rating by modality included roles and responsibilities (mean score differences of 0.4 and 0.5 points for stations 1 and 2, respectively). Paired sample t-tests revealed statistically significant differences in student station scores between in-room and video ratings, including the calculated overall average student performance score across the two stations, \( t(54) = 6.6, p < .001 \). There was a mean difference of 0.3 points between the overall average score for in-room (mean score 2.1, SD 0.4) and video (mean score 1.7, SD 0.4) ratings.

The pre-specified criterion for a passing score was achieving an average score of 2 (at expected) out of 3 (above expected) for all items except the global score, for at least one of the two stations. A series of McNemar’s test indicated significant differences in determination of pass/fail status based on scoring 2 out of 3 on average across all items between in-room and video rating for Station 1, \( p < .001 \), but not for Station 2, \( p = .027 \). There was also a significant difference in pass/fail determination of overall TOSCE performance (passing at least 1 of 2 stations), \( p < .001 \). Using the criterion of an average score of 2 out of 3 for at least one station to pass the TOSCE overall, 56% of students passed when rated in-room, compared with 20% of students who passed when rated by video (Table 2).

**DISCUSSION**

This pilot study was conducted to examine equivalence between in-room and video ratings of student team
behaviors during a TOSCE. There was no equivalence between the two modalities of assessment. The finding concurs with a previous study, which reported that in an OSCE setting with a single pharmacy student being assessed, the two modalities did not result in equivalent pass/fail decisions despite high scale reliability and intraclass correlation. Similar to what was observed in this study, the video rating was consistently lower than the in-room rating. This study’s findings contrast with studies reporting high congruence between live and video faculty ratings for procedural skills such as joint examination, airway insertion, and septic shock management. Only one pilot study of student team performance suggests reasonable internal consistency between live and video-based faculty ratings of students in teams.

There are several potential explanations for the results seen in this study. First, team behaviors are primarily based on communication skills (verbal and nonverbal) not easily captured on camera. The in-room rater was closer and had greater access to the finer nuances of communication and the connection or “chemistry” between students and the patient as well as among students, to perceive and rate those behaviors. Second, the camera angle captured only one distant perspective for the video rater, whereas the in-room rater had opportunity to change the observation perspective by moving their heads without change in the seated position, thus giving them more information on body language. Third, it is a challenge for a rater to simultaneously score several students interacting with one another and with the patient in multiple categories of behaviors. Whether in-room and video scores could be more congruent when fewer than four students are being rated at once remains to be studied. And lastly, the video rater in this study was an expert trainer, and may have applied stricter scoring to be studied. And lastly, the video rater in this study was an expert trainer, and may have applied stricter scoring to be studied. Whether in-room and video scores could be more congruent when fewer than four students are being rated at once remains to be studied. And lastly, the video rater in this study was an expert trainer, and may have applied stricter scoring to be studied.

Table 2. Student Pass/Fail Status When Pass is Determined by Average Score of 2 Out of 3 Across Competencies for at Least One of Two Stations

| Station | In-room rating | Video rating | McNemar’s Test Statistic |
|---------|----------------|--------------|-------------------------|
|         | Students Who Passed | N (%)       | Students Who Passed | N (%)       | df=1  |
| Station 1 | 63              | 34 (54)      | 55                      | 9 (16)      | 14.09* |
| Station 2 | 63              | 38 (60)      | 47                      | 16 (34)     | 4.76   |
| Overall  | 63              | 35 (56)      | 55                      | 11 (20)     | 13.14* |

*p<.01, p-value adjusted using Bonferroni correction

In their performance, a TOSCE requires assessment of individual student performance, even though each performed as a member of a team. The TOSCE challenges faculty to make these more refined judgments using the 3-point scale, while distinguishing student performance from that of team performance. Reasons aside, the differences in pass/fail decisions between the two different modalities is striking. Careful consideration should be made when determining performance assessment modality in high stakes situations.

This study has several strengths, including the use of a validated scale. Also, the same rater training was conducted for both rating modalities (in-room and video). The rigor of training is evidenced by the small, non-significant differences in scoring between the two in-room raters at each station. Visual perspective was consistent by ensuring appropriate camera placement; and the video rater viewed all stations in their entirety once to simulate the in-room viewing condition. Students earned the full range of performance scores across both stations, i.e., the scores were not uniform. Limitations of the study include having only one video rater, and having complete video ratings for only 47 of 63 students due to technical shortcomings. Because students rotate as part of a 4-person team in a TOSCE, obtaining scores for a large number of teams is difficult, requiring large numbers of students. TOSCE stations are longer than traditional OSCE stations. Thus, slightly increasing only the number of teams participating in the TOSCE can substantially increase faculty time. Because of these limitations, the number of teams participating in the TOSCE was restricted, precluding analysis of team-level performance.

Systematic reviews suggest that simulation-based teaching and assessment are superior to traditional classroom teaching for achieving specific clinical skills and improving patient care outcomes. Additionally, educators have the opportunity to use simulation in IPE assessment by implementing TOSCEs. However, the reliability of simulation-based assessment may depend not only on the choice of rating scale and training of the rater, but also on the rating modality selected. This study brings us a step toward understanding the non-equivalence of live and video ratings in
assessing team behaviors. Caution should be exercised when decisions are made about rating modality, in particular, when multiple students are simultaneously assessed in a clinical encounter. Future studies will examine larger sample sizes; the use of multiple camera angles for capturing team behaviors; and the role and accuracy of students rating their own performance with videos, compared with in-room and video faculty ratings.

CONCLUSION

In-room and video ratings of student interprofessional team performance by trained faculty raters are not equivalent. Scores based on the video ratings may reflect some limitations of the modality rather than of the student. We recommend that educators consider scoring discrepancies based on modality when assessing team behaviors.

ACKNOWLEDGMENTS

The authors are grateful to the students and faculty who participated in the project, and to Kevin Lohenry, PhD and Christopher P. Forest, PA-C for guidance and administrative support; and Anne Walsh, PA-C and Melissa Durham, PharmD for manuscript review. This project is supported by the Health Resources and Services Administration (HRSA) of the U.S. Department of Health and Human Services (HHS) under grant #D57HP23251 Physician Assistant Training in Primary Care, 2011-2016. The information or content and conclusions are those of the authors and should not be construed as the official position or policy of, nor should any endorsement be inferred by HRSA, HHS or the U.S. Government.

REFERENCES

1. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach. 2010;32(8):676-682.
2. Harden RM. Revisiting ‘Assessment of clinical competence using an objective structured clinical examination (OSCE).’ Med Educ. 2016;50(4):376-379.
3. The Medical Council of Canada. http://www.mcc.ca. Accessed March 30, 2017.
4. The United States Medical Licensing Examination. http://www.usmle.org. Accessed March 30, 2017.
5. Zorek J, Rahel C. Interprofessional education accreditation standards in the USA: a comparative analysis. J Interprof Care. 2013;27(2):123-130.
6. Solomon P, Marshall D, Boyle A, et al. Establishing face and content validity of the McMaster-Ottawa team observed structured clinical encounter (TOSCE). J Interprof Care. 2011;25(4):302-304.
7. McMaster/Ottawa TOSCE (Team Observed Structured Clinical Encounter) Toolkit. http://ifs.mcmaster.ca/tosce/en/toolkit_guidelines.html. Accessed March 30, 2017.
8. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Med Educ. 2011;45(12):1181-1189.
9. Patrício MF, Julião M, Fareleira F, Cameiro AV. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? Med Teach. 2013;35(6):503-514.
10. Simmons B, Egan-Lee E, Wagner SJ, Esdaile M, Baker L, Reeves S. Assessment of interprofessional learning: the design of an interprofessional objective structured clinical examination (iOSCE) approach. J Interprof Care. 2011;25(1):73-74.
11. Vivekananda-Schmidt P, Lewis M, Coady D, et al. Exploring the use of videotaped objective structured clinical examination in the assessment of joint examination skills of medical students. Arthritis Rheum. 2007;57(5):869-876.
12. Cook DA, Hamstra SJ, Brydges R, et al. Comparative effectiveness of instructional design features in simulation-based education: systematic review and meta-analysis. Med Teach. 2013;35(1):e867-898.
13. Lie DA, Richter-Lagha R, Forest CP, Walsh A, Lohenky K. When less is more: validating a brief scale to rate interprofessional competencies. Med Educ Online. 2017;22(1):1314751.
14. Lie D, May W, Richter-Lagha R, Forest C, Banzali Y, Loheney K. Adapting the McMaster-Ottawa scale and developing behavioral anchors for assessing performance in an interprofessional Team Observed Structured Clinical Encounter. Med Educ Online. 2015;20(1):26691.
15. Forest CP, Lie DA, Ma S. Evaluating interprofessional team performance: a faculty rater tool. MedEdPORTAL. 2016;12:10447. https://www.mededportal.org/publication/10447. Accessed April 6, 2017.
16. Sturpe DA. Objective structured clinical examinations in doctor of pharmacy programs in the United States. Am J Pharm Educ. 2010;74(8):Article 148.
17. Vivekananda-Schmidt P, Lewis M, Coady D, et al. Exploring the use of videotaped objective structured clinical examination in the assessment of joint examination skills of medical students. Arthritis Rheum. 2007;57(5):869-876.
18. House JB, Dooley-Hash S, Kowalenko T, et al. Prospective comparison of live evaluation and video review in the evaluation of operator performance in a pediatric emergency airway simulation. J Grad Med Educ. 2012;4(3):312-316.
19. Williams JB, McDonough MA, Hilliard MW, Williams AL, Cunioowski PC, Gonzalez MG. Internmethod reliability of real-time versus delayed videotaped evaluation of a high-fidelity medical simulation septic shock scenario. Acad Emerg Med. 2009;16(9):887-893.
20. Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: rethinking the etiology of rater errors. Acad Med. 2011;86(10 Suppl):S1-S7.
21. Emmert MC, Cai L. A pilot study to test the effectiveness of an innovative interprofessional education assessment strategy. J Interprof Care. 2015;29(5):451-456.
22. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. Acad Med. 2011;86(6):706-711.
23. Brydges R, Hatala R, Zendejas B, Erwin PJ, Cook DA. Linking simulation-based educational assessments and patient-related outcomes: a systematic review and meta-analysis. Acad Med. 2015;90(2):246-256.
24. Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. Acad Med. 2013;88(6):872-883.