EDUCATION
RESEARCH SERVICE

ORIGINAL ARTICLE

The interrater reliability of an objective structured practical examination in measuring the clinical reasoning ability of chiropractic students

Kevin A. Rose DC, MPH and Jesika Babajanian DC, MSHPE

Objective: The objective structured practical examination (OSPE) is a case-based assessment that can be used to assess the clinical reasoning ability of students. The reliability of using an OSPE for this purpose has not been reported in the literature. The objective of this study was to determine the interrater reliability of the OSPE in measuring the clinical reasoning ability of chiropractic students.
Methods: Two examiners tested each student simultaneously when enough were available as a check for interrater reliability. The scores for students over 4 exam administrations were compiled, and we calculated an intraclass correlation coefficient (ICC) using 1-way random single measures.
Results: Paired scores were available for 133 students. The ICC was .685, showing a fair-to-good level of agreement for faculty in assessing the clinical reasoning ability of chiropractic students using an OSPE.
Conclusion: The OSPE can be a valuable tool for testing clinical reasoning abilities because it can simulate the decision-making process that needs to be implemented in clinical practice. Faculty members at our chiropractic college were able to achieve an acceptable level of reliability in measuring the clinical reasoning abilities of students using an OSPE. Other health professional programs may consider using this tool for assessing the clinical reasoning skills of their students.

Key Indexing Terms: Decision Making; Reliability of Results; Chiropractic; Education

J Chiropr Educ 2016;30(2):99–103 DOI 10.7899/JCE-15-16

INTRODUCTION

Critical thinking and clinical reasoning are essential components of health care practice. Although these terms are sometimes used interchangeably, critical thinking is a more general term that has been described as "purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based."[1] Clinical reasoning is the application of critical thinking in clinical practice. Clinical reasoning is needed to manage complex decision making, such as integrating patient historical findings and examination results into appropriate diagnoses and deriving successful management plans in busy health care environments with incomplete data.[2] Clinical reasoning is necessary to practice evidence-informed health care, especially when there is still uncertainty regarding best practices.[3] Clinical reasoning is a core component of the assessment and diagnosis metacompetency of the Council on Chiropractic Education.[4] These metacompetencies are used to determine program learning outcomes (PLOs) by many chiropractic training programs.

Traditional health profession education programs typically focus on the development of discipline-specific knowledge and skills. More recently, educators have realized the need to also develop thinking strategies that support clinical practice.[2] There is currently limited understanding of how clinical reasoning should best be developed or objectively assessed.[5] Ramaekers et al.[6] point out that students should be given challenges similar to those they will see in their future practices to determine if they have the proper problem-solving and decision-making capabilities.

The objective structured clinical examination (OSCE) has emerged as the gold standard for the assessment of clinical competence of health professionals. It provides objectivity, structure, simulation of real clinical situations, and an assessment of various skills and competencies.[7] However, the use of stations makes the OSCE very unlike clinical practice, and it often emphasizes the performance of clinical skills over the assessment of clinical reasoning. The objective structured practical examination (OSPE) has

been described as a tool to more accurately assess the clinical reasoning abilities of health profession students.[8] The OSPE differs from the OSCE in that it is case-based instead of station-based, although there is considerable interchanging of the terms in the literature. The use of an unfolding case format allows a closer simulation of what students need to perform in practice.

The literature on OSPEs does not include tests of its reliability. The objective of this study was to determine the interrater reliability of the OSPE in measuring the clinical reasoning ability of chiropractic students.

METHODS

Description of Our OSPE

The Los Angeles College of Chiropractic, part of the Southern California University of Health Sciences, has included clinical reasoning as one of its PLOs since 2009. At that time the competency of students in clinical reasoning was largely evaluated by their primary clinical supervisor by observations during clinic shifts. In 2011, the clinical faculty and administration decided to develop an OSPE to better evaluate students' clinical reasoning skills.

Pilot tests of a preliminary version of the OSPE were conducted with students with a wide range of perceived competence in clinical skills. This testing disclosed confusing aspects of the exam for students and faculty, which were addressed in subsequent iterations. It also gave an indication of the amount of time to allow for completing the exam. It was decided to not include the demonstration of any clinical skills as this was already being assessed in a series of OSCEs. The OSPE was developed as an oral exam focusing solely on clinical reasoning. An unfolding case format was developed, which served to keep students on track during each stage of the exam. Once the format was finalized, a total of 18 cases were then created.

The OSPE was initiated in 2012. It is given to 3rd-year students during their clinical internship. They are given a practice session in a prior class, but they are not told which cases will be used. Passing this exam is mandatory for graduation. Since this is a high-stakes examination, the program dean requested that 2 faculty members test each student simultaneously when possible. This serves as a check for the reliability of scoring and as a safeguard against claims of prejudice of a faculty member against a student.

Students are given 20 minutes to complete the OSPE. A warning is given when there are 2 minutes remaining. Other than that, students are required to use their own skill in time management to determine how much time to allocate for each section.

The sequence during the OSPE is as follows:

1. Students are provided a brief history of a patient with a distinct complaint and then asked to provide 5 differential diagnoses, including 3 neuromusculoskeletal (NMS) and 2 systemic conditions, and provide substantiation for each.
2. They are then given 1 NMS and 1 systemic condition to focus on for the remainder of the exam and are asked to provide 5 history questions that can help rule in or rule out the 2 conditions. They are asked to provide the expected answer for each question for each condition.
3. Students are then asked to provide 5 physical examination procedures that can help rule in or rule out the 2 conditions, including the expected findings, and asked to provide 2 diagnostic studies that can help rule in or rule out the 2 conditions, including the expected findings.
4. They are then provided the results of selected additional history questions, physical examination procedures, and diagnostic studies and asked to provide a working diagnosis, along with their rationale.
5. Finally, students are given the working diagnosis for the case and asked to develop a report of findings for the patient, including a comprehensive explanation of the patient's condition, management plan, treatment alternatives, and risks.

The examiners are supplied with scoresheets to grade the students' performance. The complete set of scoresheets is presented in Appendix A, an online supplement that accompanies this article at www.journalchiroed. The scoresheet keeps track of student responses and reasoning and contains a list of possible appropriate answers for each section, along with space to document the examinees' substantiation. Check marks are used to keep track of the number of appropriate items provided by the student. There are extra checkboxes for the examiner to enter other answers supplied by the student. These can be counted among the correct responses if deemed appropriate by the examiner. Examiners communicate with students during the exam to ask them to elaborate on their thinking process.

Each of the 6 sections of the exam is scored on a 4-point scale using a rubric. The number of check marks on the scoresheet and the students' quality of substantiation are used as descriptions in the cells of the rubric. The examiner can also enter additional comments about students' performance for each section, such as on the organization of their thinking process.

At the end of the exam, another rubric requires the examiner to generate scores for the student in the areas of professionalism, communication, and overall clinical reasoning ability. Finally, examiners are asked for specific overall comments about the student's areas of achievement, areas needing improvement, and if they feel the student requires remediation in clinical reasoning.

After the OSPE, the scoresheets are inspected for completeness of scoring, entered into a spreadsheet, and analyzed by the OSPE coordinator. When pairs of scores for a student vary widely, their scoresheets are examined to try to determine the cause of the discrepancy. A poor score for a section is defined as 2 or less out of the 4 points possible. Students who score poorly on 1 or 2 sections of the OSPE are referred to their primary faculty supervisor for remediation. Students who are found to be weak in more than 2 sections and/or are flagged for remediation by

an examiner are referred to our clinical skills enhancement program. They receive extra training in the areas where they demonstrated a lack of competency and are required to retake the OSPE the next time it is offered.

Study Design

This study employed a retrospective design to test the interrater reliability of the OSPE. Scores of all students who were tested simultaneously by 2 examiners between spring trimester 2014 and summer trimester 2015 (4 administrations of the exam) were compiled for analysis. Approval was received from the Southern California University of Health Sciences institutional review board prior to the commencement of this research study.

Participants

Examinees were 3rd-year chiropractic students. Examiners were either Los Angeles College of Chiropractic faculty members or residents. There was a core group of clinic faculty who had administered the exam multiple times since its inception in 2012. These were designated as experienced examiners. However, as there were not enough of experienced examiners to conduct the OSPE, other faculty members and residents were recruited as needed to assist. These were designated as novice examiners.

Procedures

The examiners met before the administration of each exam to help standardize the procedures and scoring. Examiners were instructed to not communicate during the exam so as to maintain the independence of their scoring.

Examiners were stationed in 5 to 7 rooms during the exam. Two examiners were assigned to each student to the extent that there were enough available. They were assigned pseudorandomly by the OSPE coordinator. Typically, about two-thirds of the rooms had 2 examiners. Substitutions were done after every few students to give examiners breaks and to accommodate their work schedules as needed. There were typically 7 to 9 rounds of students during each exam administration, and most examiners ended up scoring exams in the same room with 2 to 3 others.

Assignment of students to examiners was done in a pseudorandom fashion by the OSPE coordinator. Students were led into the building and were positioned in front of a room in the order in which they filed in. Each 3rd-year student had a primary clinical supervisor. These supervisors were not allowed to test their own students to help prevent bias. Students were switched to other rooms as necessary to prevent this. Frequent rotation of cases and segregating students was used to minimize the effects of communication about the exam.

Analysis

Paired scores from the 4 OSPE administrations were imported into analysis software (SPSS version 23; IBM, Armonk, NY), and an intraclass correlation coefficient (ICC) was calculated. This study analyzed examination results in which 2 examiners assessed each student. Therefore, the ICC analysis was conducted using 1-way random single measures, or ICC(1,1).

Subgroup analysis was performed for different OSPE cases and examiner experience with administering the exam. Examiner experience was coded into 3 categories: 2 experienced examiners; 1 experienced and 1 novice examiner; 2 novice examiners.

Fleiss[9] states that although there are no universal standards for reliability, a good general guide is that values below .4 are considered poor, .4–.75 are considered fair to good, and above .75 is considered excellent.

RESULTS

A total of 216 students took the OSPE between spring 2014 and summer 2015. Examinations were scored by a total of 29 examiners. The average score was 77% ± 0.9%. Fourteen students (6%) failed their examination and were scheduled for remediation and a retake.

There were 133 exams in which students were simultaneously scored by 2 examiners for interrater reliability analysis. In the remaining 83 cases, only 1 examiner scored the student because of a lack of available examiners. Three cases were used by 30 students or more and were used for subgroup analysis. The remaining cases were used by 9 students or fewer in this dataset.

Table 1 shows the ICC for single measures for the entire cohort and for subgroup analyses. For the entire cohort, the reliability was in the fair-to-good range. The subgroup analysis by case showed the reliability of 2 to be in the excellent range and the 3rd in the fair-to-good range. The subgroup analysis by examiner experience showed all to be in fair-to-good range for reliability, although within this range the combination of an experienced and novice examiner testing together was the highest and 2 experienced examiners testing together was the lowest.

DISCUSSION

Reliable and valid evaluation tools are necessary to help ensure that students are meeting all the PLOs of their program. Clinical reasoning is a required component of all health profession educational programs because of the need for practitioners to be able to manage complex clinical scenarios. Unfortunately, clinical reasoning is difficult to measure because of the wide range of skills and behaviors that need to be demonstrated. Although there are validated tools that measure critical thinking, which is a part of clinical reasoning, they measure only habits of mind and not the integration of clinical skills.[3]

Ramaekers[6] described the development of the Script Concordance Test (SCT) for veterinary students, which has an unfolding case construct somewhat similar to the OSPE, albeit with several, shorter cases.[6] Their examiners achieved a reliability of .79. Goulet[10] found a SCT reliability score of .9 for practicing physicians who had been flagged as needing remediation in Canada. Selim[7] described the use of an objective structured video exam (OSVE) for psychiatric nursing students.[7] The OSVE was structured similarly to the OSPE, except video clips were

Table 1 - Intraclass Correlation Coefficient (ICC) for Single Measures for the Interrater Reliability of Scoring the OSPE

Cohort	n	ICC
Entire cohort	133	.685
By case		
Prostatic carcinoma and lumbar sprain and strain	45	.891
Pyelonephritis and lumbar facet syndrome	33	.726
Pancoast tumor and shoulder adhesive capsulitis	30	.811
By faculty experience		
2 experienced examiners	29	.562
1 experienced and 1 novice examiner	79	.745
2 novice examiners	25	.619

OSPE indicates objective structured practical examination.

shown and students in a classroom answered written questions in the areas of knowledge, observation, and clinical reasoning. They reported a reliability of .714 for their exam.

Prior to the implementation of the OSPE, our clinical faculty relied on interactions with students during clinic shifts to evaluate students' clinical reasoning skills. This process was very subjective and difficult to perform during busy clinic shifts. It was also very dependent on the particular cases that each student happened to see. Implementation of the OSPE appears to have made the evaluation of clinical reasoning more objective and reliable.

Finding that an assessment instrument is reliable is a necessary step in the path to proving its validity. This study showed an acceptable level of interrater reliability for an OSPE in the evaluation of chiropractic students' clinical reasoning ability. This is encouraging as the exam is still fairly new and each administration has involved examiners scoring it for the first time. Reliability was higher for some cases than for others, which suggests that there may be less of a consensus regarding the diagnosis and/or management of some conditions. It may be appropriate to modify or drop cases from the OSPE rotation if reliability seems lower than average.

The subgroup analysis of examiners categorized by experience in administering the OSPE showed unexpectedly that reliability was highest when an experienced examiner worked with a novice and lowest when 2 experienced examiners worked together. It is possible that 2 experienced examiners are more set in their process of how they score the exam, and these may differ from other experienced examiners. Further studies would have to be conducted to determine why this may have occurred and how to improve reliability with this combination of examiners.

Further training of examiners and/or revision of the OSPE should be performed to try to increase the reliability to a higher level. As an example of this, video vignettes of intern–patient interactions can be shown during a faculty meeting. Each examiner can be asked to first score the student separately, and then all scores can be compared to determine their level of agreement with the performance. Discussions should be held as necessary to resolve any major disagreements among participants.

Limitations

The OSPE was modified incrementally over the course of the period of study. Although the changes were small, it is possible that the reliability was higher for some versions of the exam than others. The timing of these changes was not recorded, so it was not possible to conduct a subgroup analysis of different exam versions.

Because of staffing issues, only 62% of the exams were scored simultaneously by 2 examiners. Students who were taking a retake OSPE due to poor prior performance were given priority for being scored by 2 examiners. It is possible the students with poorer clinical reasoning skills are harder to evaluate, leading to a lowering of the exam's overall reliability. It was not recorded in the dataset which students were taking a retake exam, so a subgroup analysis could not be conducted. It is known that only a small percentage of the students were taking a retake exam.

Although examiners who were scoring the OSPE in pairs were instructed not to discuss their scoring, it is possible that some communication and/or looking at the other scoresheet did occur. This OSPE was only administered to chiropractic students at 1 college. Results for other chiropractic colleges and other health profession educational programs may be different.

CONCLUSION

Faculty members at our chiropractic college were able to achieve an acceptable level of reliability in measuring the clinical reasoning abilities of students using an OSPE. Other health professional programs may consider using this tool for assessing the clinical reasoning skills of their students. Further training of examiners should be conducted to continue to improve the reliability of the OSPE. This study should be replicated with other chiropractic colleges and students of other health care professions as additional steps toward the validation of the OSPE as an evaluation tool for clinical reasoning.

FUNDING AND CONFLICTS OF INTEREST

This work was funded internally. The authors have no conflicts of interest to declare relevant to this work.

About the Authors

Kevin Rose is the university assessment coordinator at Southern California University of Health Sciences (16200 Amber Valley Drive, Whittier, CA, 90604; KevinRose@scuhs.edu). Jesika Babajanian is a clinical assistant professor at Southern California University of Health Sciences (16200 Amber Valley Drive, Whittier, CA, 90604; JesikaBabajanian@scuhs.edu). Address correspondence to Kevin Rose, 16200 Amber Valley Drive, Whittier, CA,

90604; KevinRose@scuhs.edu. This article was received July 23, 2015; revised December 11, 2015 and January 29, 2016; and accepted February 1, 2016.

Author Contributions

Concept development: KAR, JB. Design: KAR, JB. Supervision: KAR. Data collection/processing: KAR. Analysis/interpretation: KAR. Literature search: KAR, JB. Writing: KR, JB. Critical review: KR, JB.

REFERENCES

1. Facione PA. *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instructions*. Milbrae, CA: The California Academic Press; 1990.
2. Bartlett D, Cox P. Measuring change in students' critical thinking ability: implications for health care education. *J Allied Health*. 2002;31(2):64–69.
3. Carter AG, Creedy DK, Sidebotham M. Evaluation of tools used to measure critical thinking development in nursing and midwifery undergraduate students: a systematic review. *Nurse Educ Today*. 2015;35:864–874.
4. The Council on Chiropractic Education. *CCE Accreditation Standards – Principles, Processes & Requirements for Accreditation*. January 2013. http://www.cce-usa.org/uploads_2013_CCE_ACCREDITATION_STANDARDS.pdf. Accessed September 22, 2015.
5. Durning SJ, Ratcliffe T, Artino A, et al. How is clinical reasoning developed, maintained, and objectively assessed? Views from expert internists and internal medicine interns. *J Contin Educ Health Prof*. 2013;33(4):215–23.
6. Ramaekers S, Kremer W, Pilot A, van Beukelen P, van Keulen H. l: The script concordance test method. *Assess Eval High Educ*. 2010;35(6):661–673.
7. Selim A, Dawood E. Objective structured video examination in psychiatric and mental health nursing: a learning and assessment method. *J Nurs Educ*. 2015;54 (2):87–95.
8. Yaqinuddin A, Zafar M, Ikram MF, Ganquly P. What is an objective structured practical examination in anatomy? *Anat Sci Educ*. 2013;6(2):125–313.
9. Fleiss JL. *The Design and Analysis of Clinical Experiments*. Wiley Classics Library Ed. New York, NY: Wiley; 1999.
10. Goulet F, Jacques A, Gagnon R, Charlin B, Shabah A. Poorly performing physicians: Does the script concordance test detect bad clinical reasoning? *J Contin Educ Health Prof*. 2010;30(3):161–166.