Consistency of Interrater Scoring of Student Performances of Osteopathic Manipulative Treatment on COMLEX-USA Level 2-PE

Jeanne M. Sandella, DO
Larissa A. Smith, PhD
Dennis J. Dowling, DO

From the Department of Clinical Skills Testing at the National Board of Osteopathic Medical Examiners in Conshohocken, Pennsylvania.

Financial Disclosures: None reported.

Address correspondence to Jeanne M. Sandella, DO, Vice President for Clinical Skills Testing, 101 W Elm St, Suite 150, Conshohocken, PA 19428-2004.

E-mail: jsandella@nbome.org

Submitted September 27, 2013; final revision received January 2, 2014; accepted February 6, 2014.

Context: Assessment of osteopathic manipulative treatment (OMT) is included in the National Board of Osteopathic Medical Examiners' Comprehensive Osteopathic Medical Licensing Examination-USA Level 2-Performance Evaluation (COMLEX-USA Level 2-PE). The scores earned for OMT should be equivalent among all raters regardless of which technique is scored or which rater is scoring the performance. As a quality assurance measure, selected examination dates and the encounters within the administration of COMLEX-USA Level 2-PE are scored by 2 raters: first by a "live" rater and next by a quality assurance rater. Neither rater knows if he or she is the first or second rater.

Objective: To compare candidate's scores recorded for OMT on COMLEX-USA Level 2-PE to determine whether differences exist among raters and techniques scored.

Methods: The authors evaluated candidate performances that took place from July through November 2012. For each performance, 2 raters scored the same technique or different techniques using the OMT scoring rubric. Discrepancies between scores were compared using *t* tests. Statistical significance was set at $P<.05$ for most analyses.

Results: Of the 708 performances, there was no statistically significant difference in scoring whether the OMT raters scored the same technique or different techniques when the students performed more than 1. There were no statistically significant differences between these results and instances when only a single technique was performed and scored.

Conclusion: The present study provides reliability evidence for the use of the global OMT scoring tool in the evaluation of OMT in COMLEX-USA Level 2-PE.

J Am Osteopath Assoc. 2014;114(4):253-258
doi:10.7556/jaoa.2014.050

High-stakes assessment of clinical skills is now requisite for all physicians in the United States.[1-3] Not only are these examinations designed to assess history taking and physical examination skills, but also they may evaluate a candidate's technical abilities through the use of simulators and other technologies to assess the performance of complex multistep processes.[4-7] Candidates' skills may be gauged by means of direct observation, objective structured clinical examination, and oral examination. Each method has its advantages, and the choice of which assessment depends, in part, on the purpose of the assessment. Is the examination primarily an educational tool? Is the examination conducted to identify candidates in need of remediation or to make decisions on advancement, promotion, or licensure?

Examinations that are used for high-stakes decisions—in such areas as advancement, promotion, or licensure—must be designed carefully to ensure accurate and reliable performance measures. Such assessments must take into account a complicating factor: during performance assessments, candidates may realize multiple correct ways of reaching an objective. The evaluation of a candidate's clinical skills can involve various types of rating tools, from checklists to key action item analyses linked with timing and global rating scales.[4,6,8-11]

The National Board of Osteopathic Medical Examiners (NBOME) administers the Comprehensive Osteopathic Medical Licensing Examination-USA Level 2-Performance Evaluation (COMLEX-USA Level 2-PE). The scoring of the Biomedical/Biomechanical Domain involves evaluation of skills in history taking, physical examination, documentation of the encounter in a SOAP (subjective, objective, assessment, plan) note, and performance of osteopathic manipulative treatment (OMT).[1,12,13] Trained osteopathic physician–examiners (ie, raters) evaluate the 25% to 40% of the candidate's encounters that are specifically scored for OMT performance. The raters score each encounter using a global Likert-scale evaluation tool, which was developed by content experts for the NBOME to provide the means to assess OMT in COMLEX-USA Level 2-PE. This tool is applied to all encounters in which OMT is scored, regardless of the content. A rigorous review is conducted on an ongoing basis to establish interrater reliability by means of statistical analyses (including quality assurance [QA]) and double-scoring encounters. Raters for the OMT portion of the examination are trained and experienced in scoring student-performed OMT. They are able to recognize variations on techniques as well as the orthodox means of executing the treatment. All OMT raters must participate in yearly refresher training.

During COMLEX-USA Level 2-PE, candidates are permitted to perform any type of OMT technique (except for high-velocity, low-amplitude and other articulatory thrust techniques) on each standardized patient (SP). It is important to ensure that the scores earned for this skill are reproducible within a given encounter. Candidates may perform multiple techniques (eg, soft tissue, then facilitated positional release, then muscle energy), 1 technique, or none at all. The scoring rubric, however, is applied to the 1 technique that a rater regards as the best performed by the candidate in a single SP encounter. The rater documents this technique and also records the other techniques that are observed but not scored. But what if 1 of the nonscored techniques was scored instead? How, if at all, would such a shift change a candidate's overall score? In the present study, we examined how scores would compare if the 2 raters scored 2 different techniques performed by the same candidate. Would these scores be as accurate as the scores earned by a candidate who performs only 1 technique—a situation that requires both raters to score the same technique?

We hypothesized that a candidate's scores would not differ regardless of rater or technique scored.

Methods

Institutional review board approval for this study was obtained through the Center for the Advancement of Healthcare Education and Delivery.

Examination

The COMLEX-USA Level 2-PE is a competency-based pass-fail examination of clinical skills. The present study followed examination protocol: candidates rotated through 12 SP stations and evaluated and treated SPs as they saw fit during each 14-minute encounter. Standardized patients simulated various conditions—such as gait changes or limitation of motion and tenderness—and responded to questions in a scripted, consistent manner. Candidates were not told which encounters were predetermined to be scored for OMT. They were told to use their clinical judgment to decide which SP would benefit from osteopathic evaluation or from OMT during the encounter. Candidates were instructed to limit the OMT performed to 3 to 5 minutes and informed that they were not required to treat an SP to a desired clinical endpoint (eg, complete resolution of symptoms). High-velocity, low-amplitude and articulatory thrust techniques were prohibited in the examination.

Scoring

All candidate-SP encounters were digitally video recorded. Two cameras with pan-tilt-zoom function—controlled by operators in a separate room—were strategically placed at 90° angles facing inward to capture the optimum views of the candidate-SP encounter. The OMT performance was scored by osteopathic physicians who underwent training as examiners. Examiners signed on to a secure Web-based portal where they had access to assigned candidate videos from the examination. They were provided case summaries and were specifically assigned to cases and therefore became familiar with the case materials and details.

The scoring rubric was developed by osteopathic physicians for the purpose of evaluating the performance of OMT in COMLEX-USA Level 2-PE.[1,5,12,13] The rubric is reviewed annually by a subcommittee of osteopathic physicians to make sure it remains clear and current. It consists of the following 6 dimensions: Osteopathic Examination/Evaluation, Patient/Physician Position for Treatment, OMT Modality Selected, OMT Technique, Treatment Repetition/Duration, and Post-Treatment Assessment. Examiners score each of 6 dimensions on a 9-point Likert scale that is broken into 3 performance groups: Unacceptable, Standard, and Superior. These ranking groups have descriptive statements, which clearly describe each level of behavior across the continuum.

In addition, the OMT raters watched each video recording and used his or her judgment to identify which technique or techniques were being performed by the candidates during their encounters. Using the *Glossary of Osteopathic Terminology*,[14] a predetermined list of techniques was developed by physician staff of the NBOME. The technique list was finalized and approved by a consensus of experienced OMT raters.[13] The website portal scoring rubric allows for OMT raters to select multiple techniques being performed by a candidate, but only 1 technique may be selected for scoring OMT. To reflect the candidates' capabilities, only the better or best performed OMT intervention—as judged by the rater—is scored in its entirety.

Sample

Data were taken from candidates who took COMLEX-USA Level 2-PE between July 12, 2012, and November 28, 2012. The data set contained all candidates who were double-scored as part of routine QA procedures.

The representative sample was derived from 28 colleges of osteopathic medicine and included a candidate's ethnicity, sex, primary language, and number of times a candidate had taken COMLEX-USA Level 2-PE.

Quality Assurance

Each OMT rater's score is subject to several QA checks throughout the testing cycle. One type of QA review entails randomly selected double scoring of encounters. Double scoring occurs when a second OMT rater—the QA rater—scores an encounter for a specific testing session that has already been scored by a first OMT rater (ie, the live rater). The QA OMT raters score the encounters using the same method as the live OMT raters: review of the encounter video recording. All raters are used for both live and QA ratings throughout the test cycle. Because QA raters score encounters in the same manner that they would if rating as a live rater, they are unaware that they are providing a QA rating. The live ratings are used for scoring purposes, and the QA ratings are used for comparison against the live ratings and technique selections. These rating comparisons are reviewed on a monthly basis to assess interrater reliability and to identify any potential discrepancies in physician-examiner scoring and technique selection.

Statistical Analysis

We used *t* tests to compare signed and absolute discrepancy between live and QA scores. Statistical significance was set to $P<.05$ for most analyses. Descriptive statistics are provided in the Results section.

Results

Our data set represented the 708 of 2211 candidates (32%) who took the examination during the study period. The population was primarily white (65%), male (52%), and first-time COMLEX-USA Level 2-PE takers (95%), and almost all (95%) spoke English as a first language.

Raters' data were compared to see if they had scored techniques in the same category (eg, myofascial technique) or different categories (eg, rib-raising vs sinus draining) for a given encounter. The current sample represented ratings data from 30 unique raters and 44 rater dyads. Each dyad rated between 12 and 48 encounters, with a mean (standard deviation [SD]) of 16.47 (7.84) encounters. If both raters scored the same technique, the ratings were coded as *matched*; if raters did not score the same technique, the ratings were coded as *mismatched*. Of the 708 rater dyads in the sample, 493 (70%) were matched and 215 (30%) were mismatched.

There were no statistically significant differences in total score between the matched and the mismatched groups ($t_{630}=1.74$). Candidates performed between 1 and 4 OMT techniques per encounter, with a mean (SD) of 1.4 (.59); the mismatched group performed significantly more techniques than the matched group ($t_{630}=-5.84$, $P<.01$), with a difference amounting to about one-third of an OMT technique ($M_{match}=1.277$, $M_{mismatch}=1.565$).

The matched and mismatched groups did not differ by sex ($\chi^2_1=1.07$), use of English as a primary language ($\chi^2_1=.10$), or ethnicity ($\chi^2_5=4.00$). It should be noted, however, that demographic data were available for approximately 75% of the sample.

Outliers, in which live vs QA dyad scores were so discrepant as to be unrepresentative of the population of rater dyads in general, were eliminated from the data set for both matched and mismatched groups by computing *z* scores for the absolute difference in rater scores and eliminating rating pairs ($z>2.33$; $P<.01$). This elimination resulted in 8 data points being dropped from the matched group and 10 from the mismatched group, leaving final sample sizes of 485 and 205, respectively.

Once the rater dyads were coded as matched or mismatched, the absolute and signed differences between their ratings were computed. The authors then performed *t* tests to assess whether rater agreement was lower when raters scored different techniques than when they scored the same ones.

The mean (SD) signed differences in live and QA ratings on the 10-point scale were .0774 (.9148) for the matched dyads and .7251 (.5622) for the mismatched dyads, a statistically nonsignificant difference of less than three-fourths of a rating scale point ($t_{688}=.12$). When the means of the absolute values of the differences were examined—eliminating rater order effects and looking solely at the degree to which the 2 raters differed—the mean (SD) unsigned differences were .0679 (.9895) and .8124 (.5660), a difference of about three-fourths of a scale point, which was not statistically significant ($t_{688}=-1.86$). Live and QA raters scoring different techniques on the same candidate were no more discrepant in their ratings than live and QA raters scoring the same technique.

Because of the statistically significant difference in the number of techniques performed between the matched and mismatched groups, the analyses were rerun on the subsample of examinees who were coded by the live rater as having performed multiple techniques. With the single-technique examinees removed, the difference between the matched and mismatched groups in the technique count was no longer statistically significant ($t_{191}=-1.23$). The mean (SD) number of techniques performed in this subsample was 2.13 (.37) for the matched group and 2.20 (.43) for the mismatched group.

With the single-technique candidates eliminated, there were still no statistically significant differences in the degree of rater discrepancy in either the signed differences (t_{191}=.42) or the unsigned differences (t_{191}=.19). Even reinclusion of the outliers in the analysis did not change the results. Whether one examines the full candidate sample or only the sample originally coded as performing multiple techniques, 2 raters scoring 2 different techniques are no more discrepant than 2 raters scoring the same technique.

Discussion

In providing a means to evaluate a clinical encounter and candidate performance, it is important to establish a protocol to make judgments on the proficiency of the candidate. With OMT, there are instances in which numerous treatment methods can be used to manage a particular complaint or somatic dysfunction.

The analyses of the overall mismatched ratings by OMT technique upheld the null hypothesis: there was no difference in scoring whether the OMT raters scored the same technique or different techniques when the students performed more than 1. There were no statistically significant differences between these results and instances in which only 1 technique was performed and scored.

Candidates are instructed multiple times—including during preexamination orientation and within online materials regarding the COMLEX-USA Level 2-PE—to evaluate and treat the SPs as they see fit. The guidelines state that when a student determines that OMT is clinically appropriate after taking the patient's history and performing a focused physical examination of the patient, the student is to treat the patient for 3 to 5 minutes and does not need to treat to clinical conclusion. Therefore, they can perform 1 technique or several. Also, the other aspects of the encounter are not scored by the OMT raters, including history-taking, physical examination, and documentation.

As noted in a previous study by Langenau et al,[13] candidates taking COMLEX-USA Level 2-PE use a wide variety of OMT techniques. Generally, it could be assumed that a candidate would choose those interventions with which they were most familiar and comfortable. In a high-stakes examination that is required to be passed to graduate from osteopathic medical school and enter into graduate medical education, it would be highly inadvisable to attempt techniques that are unfamiliar or rarely used. Factors that would influence a candidate's choice of OMT technique may include initial interest and skill in OMT during the first few years of osteopathic medical school, exposure to OMT during rotations and organized training, or a mentor's advice. Depending on how these factors intermingle, a candidate may be extremely or minimally capable at performing OMT. Students' attitudes regarding OMT, as well as exposure to osteopathic medicine prior to entering osteopathic medical school, also appear to indicate skill in the performance of OMT.[15-18]

In any case, a candidate's performance of OMT should be consistent, regardless of the technique. Raters of the OMT portion of this examination are also required to record all techniques that are demonstrated. The present study shows that whether a student performs 1 technique or multiple techniques, the QA scores do not significantly vary using this rubric and are reproducible by a second, independent rater. In addition, this consistency holds true even if the raters chose different techniques to score, suggesting that the candidates' abilities are the same regardless of rater or technique. Further research on this topic across encounters would be interesting. Such research is important to help show the reliability of the raters' scores in the assessment of the candidates' Biomedical/Biomechanical Domain scores.

The limitations to the present study were mainly temporal. We were limited to a subset of the candidates examined in 4 months of the 2012 test cycle. In addition, as previously stated, it would be interesting to examine how well candidates perform OMT, as well as the other skills assessed throughout the test day and across different encounters.

Conclusion

The present study shows that students receive equivalent scores from OMT raters in COMLEX-USA Level 2-PE whether they perform 1 or more than 1 technique and regardless of which technique a rater should select to score. This finding provides additional reliable evidence for the use of the Global OMT scoring tool in the evaluation of OMT in COMLEX-USA Level 2-PE.

References

1. Langenau EE, Dyer C, Roberts WL, Wilson C, Gimpel J. Five-year summary of COMLEX-USA Level 2-PE examinee performance and survey data. *J Am Osteopath Assoc*. 2010;110(3):114-125.
2. Boulet JR, Smee SM, Dillon GF, Gimpel JR. The use of standardized patient assessments for certification and licensure decisions. *Simul Healthc*. 2010;4(1):35-42. doi:10.1097/SIH.0b013e318182fc6c.
3. Whelan GP, Boulet JR, McKinley DW, et al. Scoring standardized patient examinations: lessons learned from the development and administration of the ECFMG Clinical Skills Assessment (CSA). *Medical Teach*. 2005;27(3):200-206.
4. Mudumbai SC, Gaba DM, Boulet JR, Howard SK, Davies MF. External validation of simulation-based assessments with other performance measures of third-year anesthesiology residents. *Simul Healthc*. 2012;7(2):73-80. doi:10.1097/SIH.0b013e31823d018a.
5. Boulet JR, Gimpel JR, Dowling DJ, Finley M. Assessing the ability of medical students to perform osteopathic manipulative treatment techniques. *J Am Osteopath Assoc*. 2004;104(5):203-211.
6. Murray DJ, Boulet JR, Kras JF, Woodhouse JA, Cox T, McAllister JD. Acute care skills in anesthesia practice: a simulation-based resident performance assessment. *Anesthesiology*. 2004;101(5):1084-1095.
7. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation |and assessment of clinical skills of medical trainees: a systematic review. *JAMA*. 2009;302(12):1316-1326. doi:10.1001/jama.2009.1365.
8. Boulet JR, van Zanten M, de Champlain A, Hawkins RE, Peitzman SJ. Checklist content on a standardized patient assessment: an ex post facto review [published online July 27, 2006]. *Adv Health Sci Educ Theory Pract*. 2008;13(1):59-69.
9. Boulet JR, Smee SM, Dillon GF, Gimpel JR. The use of standardized patient assessments for certification and licensure decisions. *Simul Healthc*. 2009;4(1):35-42. doi:10.1097/SIH.0b013e318182fc6c.
10. Nicholson P, Gillis S, Dunning AM. The use of scoring rubrics to determine clinical performance in the operating suite [published online August 27, 2008]. *Nurse Educ Today*. 2009;29(1):73-82. doi:10.1016/j.nedt.2008.06.011.
11. Mudumbai SC, Gaba DM, Boulet J, Howard SK, Davies MF. Feasibility of an internet-based global ranking instrument. *J Grad Med Educ*. 2011;3(1):67-74. doi:10.4300/JGME-D-10-00162.1.
12. Gimpel JR, Boulet DO, Errichetti AM. Evaluating the clinical skills of osteopathic medical students. *J Am Osteopath Assoc*. 2003;103(6):267-279.
13. Langenau EE, Dowling DJ, Dyer C, Roberts WL. Frequency of specific osteopathic manipulative treatment modalities used by candidates while taking COMLEX-USA Level 2-PE. *J Am Osteopath Assoc*. 2012;112(8):509-513.
14. Educational Council on Osteopathic Principles. *Glossary of Osteopathic Terminology*. Chevy Chase, MD: American Association of Colleges of Osteopathic Medicine; 2011. http://www.aacom.org/resources/bookstore/Documents/GOT2011ed.pdf. Accessed February 2, 2014.
15. Johnson SM, Kurtz ME. Perceptions of philosophic and practice differences between US osteopathic physicians and their allopathic counterparts. *Soc Sci Med*. 2002:55(12):2141-2148.
16. Johnson SM, Kurtz ME, Kurtz JC. Variables influencing the use of osteopathic treatment in family practice. *J Am Osteopath Assoc*. 1997;97(2):80-87.
17. Johnson SM, Kurtz ME. Diminished use of osteopathic manipulative treatment and its impact on the uniqueness of the osteopathic profession. *Acad Med*. 2001;76(8):821-828.
18. Chamberlain NR, Yates HA. A prospective study of osteopathic medical students' attitudes toward use of osteopathic manipulative treatment in caring for patients. *J Am Osteopath Assoc*. 2003;103(10):470-478.