JAOA ORIGINAL CONTRIBUTION

Evaluating the Clinical Skills of Osteopathic Medical Students

John R. Gimpel, DO
John R. Boulet, PhD
Anthony M. Errichetti, PhD

Because clinical skills play an important role in health services, many medical credentialing organizations are making performance-based assessments part of the board-certification and licensure processes. While clinical skills are taught and evaluated at colleges of osteopathic medicine, the development and validation of standardized assessment methodologies is far from complete. The purpose of this study was to gather data to support the use of a performance-based assessment of osteopathic clinical skills. A sample of 121 fourth-year osteopathic medical students was tested using the Comprehensive Osteopathic Medical Licensing Examination–USA performance-based clinical skills examination (COMLEX–USA–PE) prototype, a standardized patient performance evaluation that involves a series of 12 simulated encounters.

Students were evaluated in a number of domains that included history taking, physical examination, osteopathic manipulative treatment techniques, written communication and clinical problem solving, and physician-patient communication. The analysis of data from 1452 standardized patient encounters suggests that reliable and valid scores can be obtained using the current prototype. The use of COMLEX–USA–PE to assess the readiness of osteopathic medical students to provide patient care in supervised graduate medical education training programs is supported.

The ability to gather patient data, perform relevant physical examinations, communicate effectively with patients and other health care professionals, and provide treatment are all important aspects of being a physician. Unfortunately, many of these skills are difficult to measure using traditional multiple-choice or other "paper-and-pencil" tests. While selected-response examinations are useful for measuring medical knowledge and clinical reasoning, they are generally not appropriate for assessing motor skills, verbal communication, or hands-on treatment. For these and other reasons, several credentialing and licensing organizations (eg, Medical Council of Canada, Educational Commission for Foreign Medical Graduates, National Board of Medical Examiners) have implemented or are planning to implement some form of clinical skills examination for medical school graduates.[1-3]

The use of objective structured clinical examinations or other standardized patient performance-based assessments is widespread.[4-7] These evaluations, often involving a series of simulated patient interactions, have been shown to provide scores with adequate psychometric properties.[8-10] However, given the wide array of potential sources of measurement error and the limited number of tasks that an individual can be asked to complete, the reliability of the scores from these types of assessments is often limited. Therefore, special care must be taken in the test development process to ensure that the examination content (eg, cases, clinical encounters) is sound and that the evaluation tools (eg, checklists, rating scales) are well defined and appropriate. In addition, once the assessment is developed, the accuracy and consistency of patient portrayal and scoring must be monitored. These measures will help to ensure that inferences based on assessment scores are valid and reproducible.

Numerous performance-based evaluations, including standardized patient assessments, have been developed to measure the clinical skills of osteopathic and allopathic medical students, residents, and physicians.[11-13] These evaluations, while primarily focusing on teaching and formative assessment, have also been used for summative purposes. In the United States, most allopathic medical schools have some form of standardized patient program. Likewise, many osteopathic medical programs use objective structured clinical examinations for training and evaluation.[14,15] While osteopathic and allopathic assessments share many of the same measurement domains, osteopathic medicine is based on a concept that the normal body, when in physiologic homeostasis, is a vital machine capable of making its own remedies against disease.

The role of the neuromusculoskeletal system in health and disease, an enhanced focus on health promotion and disease prevention, and an emphasis on the physician-patient

From the National Board of Osteopathic Medical Examiners in Chicago, Illinois, where Dr Gimpel is director of performance testing and Drs Boulet and Errichetti are consultants.

Address correspondence to John R. Gimpel, DO, Director of Predoctoral Education, Georgetown University School of Medicine, 208 Kober Cogan Hall, 3750 Reservoir Road, NW, Washington, DC 20007.

E-mail: jrg34@georgetown.edu

relationship are fundamental to the osteopathic philosophy of health care. As a result, required medical knowledge may be different,[16] with patient interactions often involving a unique set of conditions and treatment modalities, eg, osteopathic manipulative treatment (OMT). For an assessment to be valid, the individuality of the profession and associated practice patterns must be represented in the content domain. Therefore, the evaluation and assessment of the clinical skills of osteopathic physicians, as opposed to their allopathic counterparts, require additional performance measures, different patient scenarios, and tailored scoring rubrics.

The National Board of Osteopathic Medical Examiners (NBOME) administers the Comprehensive Osteopathic Medical Licensing Examination–USA (COMLEX–USA). This set of paper-and-pencil examinations is the standard for testing osteopathic medical students[17]and generally focuses on assessing the medical knowledge and clinical reasoning of osteopathic students and graduates. Unfortunately, the direct assessment of clinical skills is not currently included in the examination sequence leading to a license to practice osteopathic medicine. To address this concern, the NBOME is in the process of developing and validating a performance-based clinical skills examination.[18] This examination, known as COMLEX–USA–PE, is being designed to evaluate the clinical skills of osteopathic medical school graduates who wish to enter graduate medical education training programs. Like many other clinical skills assessments, COMLEX–USA–PE will use standardized patients—lay people trained to realistically portray patients with specific clinical problems. However, given the differences in philosophy and training between practitioners who use the diagnostic and therapeutic measures of allopathic medicine and those who also incorporate OMT and other features of osteopathic principles and practice, the nature and focus of a clinical skills assessment targeted at osteopathic physicians needs to be somewhat different than one developed specifically for allopathic physicians.

Purpose

The current investigation, pilot study A (PSA), was designed to investigate the administrative logistics and psychometrics of the COMLEX–USA–PE prototype, incorporating some revisions that were made based on data from an earlier feasibility study. These modifications included the standardization of station timing and postencounter formats, the revised role of the osteopathic physician examiner, enhancements in data collection and real-time quality assurance, augmentations to standardized patient and osteopathic physician examiner training protocols, and a lengthening of the assessment to 12 clinical encounters. The NBOME's objective in PSA was to gather data to support the validity and reliability of COMLEX–USA–PE scores for the purposes of assessing the readiness of fourth-year osteopathic medical students for entry into osteopathic graduate medical education training programs.

Methods

Pilot Study A

Pilot study A was conducted at Western University of Health Sciences College of Osteopathic Medicine of the Pacific (COMP). One hundred twenty-one fourth-year osteopathic medical students were tested using the performance-based COMLEX–USA–PE prototype examination. Eleven test sessions took place from October 6, 2001, to November 4, 2001. More than 1450 individual patient encounters were completed as part of this study.

COMLEX–USA–PE Prototype

The COMLEX–USA–PE prototype is a 12-station clinical skills examination designed specifically to assess osteopathic medical students. Standardized patients—individuals who are trained to consistently and accurately portray patients with specific medical complaints—are used for the assessment of clinical skills. The following skills are measured via COMLEX–USA–PE: medical history taking, physical examination, written communication and clinical problem solving (subjective, objective, assessment, plan; SOAP note), and physician-patient communication and relationship (global patient assessment). Also, unlike other clinical skills assessments (eg, Educational Commission for Foreign Medical Graduates/Clinical Skills Assessment; National Board of Medical Examiners standardized patient examination), there is an osteopathic emphasis, including the evaluation of OMT techniques.

The COMLEX–USA–PE assessment consists of 12 standardized patient encounters. Candidates are given 13 minutes to interview and evaluate the patient. A 7-minute postencounter exercise (written SOAP note) follows each patient interview. The set of encounters, or cases, is chosen with reference to the COMLEX–USA blueprint. As a result, candidates encounter standardized patients with a variety of complaints or reasons for visiting (reasons for visit classification) the physician. Cases are generally based on high-prevalence osteopathic medicine–specific reasons for visit classifications, but are designed to lead to a number of diagnostic outcomes. In addition, the test form was built with a set of cases that attempts to reflect patient characteristics in the general population (eg, gender, age, ethnicity, acuity of complaint).

Participants

Student Sample—One hundred nineteen students from COMP participated in PSA. In addition, two students from Touro University College of Osteopathic Medicine took the prototype COMLEX–USA–PE assessment. All of the fourth-year osteopathic medical students at COMP were required to take and successfully pass either the PSA prototype or a shortened version before graduating. To enhance motivation, students were offered prizes based on their performance, both overall and by test session. Study protocols as well as informed consent forms and releases were reviewed by the institutional review board at COMP, and the study was given exempt status.

Basic demographic data were available for the 119 students attending COMP. Mean age was 29.8 years (SD, 4.6; minimum, 24.8; maximum, 47.6). The largest ethnic group was Caucasian (n = 52; 43.7%), followed by Asian (n = 36; 30.3%) and Latino (n = 10; 8.4%). The remaining students (n = 21; 17.7%) were classified as "other." The PSA student sample was 50.4% women (n = 60) and 49.6% men (n = 59).

Standardized Patients—Twenty-four standardized patients were recruited and trained for PSA. This sample included 16 women and 8 men. For the 12-station examination, there were 8 cases portrayed by women and 4 cases portrayed by men. Mean age of the standardized patients was 52.7 years (minimum, 20; maximum, 74). The standardized patients were predominantly Caucasian (91.7%). Most of the individuals recruited for the study had some previous experience as standardized patients (n = 14; 58%).

Osteopathic Physician Examiners—Sixteen osteopathic physician examiners were used in this study for the assessment of OMT. Each osteopathic physician rater provided scores for at least one test session. The minimum number of students assessed by a given osteopathic physician rater was 11 (1 test session). The maximum number of students assessed by any given rater was 51 (5 test sessions). The specialties of the osteopathic physician examiners used in this study were family medicine, 7; internal medicine, 7 (including 2 gastroenterologists, 1 women's health specialist, 1 geriatrician); pediatrics, 1; and general surgery, 1. Full-time osteopathic manipulative medicine faculty members were excluded from this study. All examiners were board-certified osteopathic physicians with at least 3 years of clinical practice experience.

SOAP Note Raters—Four osteopathic physician examiners provided SOAP note ratings. There were three family physicians and one internist, all of whom were involved in medical education as well as clinical practice. The osteopathic physician raters had previously participated in the development of case materials for COMLEX–USA–PE and were familiar with the purpose and composition of the assessment prototype.

Training

Standardized Patients—Recruiting and training of standardized patients at COMP were conducted under the direction of NBOME staff. A 2-day workshop provided an outline and timeline of required training activities. These included the study of standardized patient training notes, training videotape review, case role-play, and documentation of standardized patient activities in detailed training logs.

Standardized patients were trained to (1) portray a patient accurately and consistently, (2) document candidate performance on the appropriate clinical skills checklists, and (3) complete the global patient assessment of doctor-patient communication skills. Two standardized patients were trained for each case (12 cases; 24 standardized patients). To assure accuracy and consistency during the examination, standardized patients playing the same case were trained together for a minimum of 8 hours by the same trainer. Enhanced standardized patient training notes, including indexed checklist items as well as training videotapes and benchmark videotapes, were used for instruction.

Osteopathic Physician Examiners—Osteopathic physician examiners were allowed to assess OMT techniques after completing 4 hours of formal training using videotapes, CD-ROMs, and hands-on demonstrations. Additionally, on each examination day, the examiners participated in a 1-hour orientation session.

SOAP Note Raters—Rater training included a 3-hour orientation session and an additional hour of case-specific videotape review, including the composition of notes. The rating guidelines, which encompass the person-centered and biomechanical aspects emphasized in osteopathic principles and practice, including standard expectations for documentation of each of the measured components, were thoroughly explained. Following training, each examiner was required to rate case-specific benchmark notes and provide data that were sufficiently accurate. In total, each osteopathic physician examiner involved in SOAP note scoring for PSA underwent 12 to 15 hours of training.

Assessment Form

A 12-case, content-balanced form of COMLEX–USA–PE was administered in PSA. The list of cases and associated patient problems is provided in *Table 1*. All students completed the same 12 cases. However, depending on the examination session, students did not encounter the same set of standardized patients. Patient interviews and treatment (where applicable) were limited to 13 minutes. The students were given 7 minutes to complete their written summaries (SOAP notes).

Scoring

History Taking—History taking was measured in each station. Case-specific checklists completed by the standardized patient after each encounter were used for scoring. These checklists consist of the relevant patient history questions that should be asked given the nature of the case and the primary patient complaint.

Physical Examination—Physical examination skills were measured in 11 (of 12) of the stations. Case-specific checklists completed by standardized patients after the encounters were used for scoring. These checklist items reflect the maneuvers that a student should complete in doing a focused physical examination. To obtain credit, students were not only required to perform specific maneuvers, but also to do them according to defined standards. As part of PSA, physical examination

checklists were also completed by osteopathic physicians in three of the stations. These identical physical examination checklists were completed for research purposes only and were not used to generate final candidate scores.

Data Gathering—The data gathering score for a given station was the percentage of history taking and physical examination items attained. The relative weighting of the history taking and physical examination components varied by case. This weighting, which is logically related to the patient problem, is a function of the number of case-specific items for each of these two skill areas as deemed appropriate by content experts.

Osteopathic Manipulative Treatment—As part of PSA, OMT was assessed in 3 (25%) of 12 of the stations. The evaluations were done by an osteopathic physician in the examination room using the recently developed OMT assessment tool. This instrument has 15 items that can each be scored from 2 (done proficiently) to 0 (done incorrectly or not done). Scores of 1 are given for actions that are done with hesitation, uncertainty, tentativeness, etc. Items assessed in this instrument were designed with input from the Educational Council on Osteopathic Principles and included conventional observable treatment behaviors, such as patient position, physician position, appropriate duration/timing, appropriate hand and finger placement, etc. A candidate's total score for a given case can range from 0 to 30. For reporting purposes and to derive an osteopathic clinical skills composite score (see next paragraph), the total can be converted to a percentage (total OMT score ÷ 30 × 100).

Osteopathic Clinical Skills—For stations where OMT is assessed, the osteopathic clinical skills score is the average of the data gathering and OMT (converted to a percentage) scores. For stations where OMT is not assessed, the osteopathic clinical skills score is simply the data gathering score.

Written Communication (SOAP Note)—While there are many ways for health professionals, including physicians, to document the information gathered during patient encounters, use of the SOAP format is common. Within this framework, physicians document what the patient told them (chief complaint, history of present illness, past medical history), what they saw in the examination (significant positive and negative physical findings), the assessment (problem list, diagnoses), and the plan (treatment, further diagnostic tests). For COMLEX–USA–PE, the notes are scored for each category (S, O, A, and P) and globally by trained osteopathic physician raters. Each note was scored for the subjective, objective, assessment, and plan portions on a 1 to 9 scale, with 1 to 3 being unacceptable and 7 to 9 being superior. Ratings of 4 to 6 were not labeled but could be considered to represent performance that was better than unacceptable, yet less than superior. The mean SOAP score (ranging from 1 to 9) is the average of the four category ratings. A global score was also given for each note, representing synthesis of the above components and accounting for legibility and overall accuracy. Values could range from 1 to 9. Written communication scores, either mean or global, can be converted to a percent metric: (SOAP-1) ÷ 8 × 100.

Table 1
Pilot Study A Case Mix

Case	Patient Problem
A	Adolescent for scoliosis/health check
B	Low back pain
C	Acute chest pain
D	Chronic abdominal pain
E	Insomnia/depression
F	Infant with gastrointestinal reflux
G	Frozen shoulder
H	Joint pain/fatigue
I	Asthmatic with cough
J	Acute dyspnea
K	Elderly patient with confusion
L	Shortness of breath/chest pain

Examination Scoring: Clinical Skills Domains—COMLEX–USA–PE is designed to assess the skill areas detailed above in two separate composites, or domains, as described.

Biomedical/Biomechanical Domain—The biomedical/biomechanical domain encompasses osteopathic clinical skills and written communication (SOAP). The biomedical/biomechanical domain score is derived by adding two thirds of the osteopathic clinical skills score to one third of the SOAP score. Values, on a percent score metric, can range from 0 to 100.

Humanistic Domain (Global Patient Assessment)—The humanistic domain includes physician-patient communication and physician-patient relationship skills. These skills are evaluated by the standardized patients in each station. The standardized patients use the global patient assessment instrument to rate the candidates across six relevant dimensions (clarity of questions, listening, explanation and summarization of information, respectfulness, empathy, and professionalism). Each dimension is rated on a scale ranging from 1 to 9 (1 to 3, unacceptable performance; 7 to 9, superior performance). The global patient assessment score for a given station is simply the mean of the six dimension scores.

Total Scores—Student scores for each composite (biomed-

ical/biomechanical, humanistic) are calculated as the average score over the 12 encounters. A student can compensate for poor performance in one encounter with excellent performance in another.

Quality Assurance

A quality assurance plan was established to ensure that standardized patients performed their respective cases and completed clinical skills checklists accurately and consistently throughout the course of the study. A 14-point performance fidelity checklist was developed and used to assess the accuracy of information exchanges between standardized patient and examination candidate as well as standardized patient body language, positioning and affect. Standardized patient trainers monitored standardized patient performance during PSA using the performance fidelity checklist. Based on these data, trainers could make adjustments to a standardized patient's performance where necessary. Checklist accuracy was examined by comparing the consistency between standardized patient checklist ratings and secondary observers rating the same student performances.

Postexamination Surveys

Immediately after each examination date, written postexamination surveys were distributed to all of the students, osteopathic physician examiners, and standardized patients. The surveys were designed to poll each respective group regarding various elements of COMLEX–USA–PE. These included questions regarding the timing for the stations and the postencounter exercise, the verisimilitude of the cases, fatigue, problems with physical examination or OMT maneuvers for standardized patients, and scoring issues for osteopathic physician examiners and standardized patients.

Analyses

Various analyses were done to investigate the psychometric adequacy of COMLEX–USA–PE scores. The reliability of COMLEX–USA–PE scores was investigated using generalizability theory.[19] Correlations and mean comparisons were used to explore the associations of COMLEX–USA–PE scores with other measures. Descriptive statistics and qualitative summaries were used to examine the assessment logistics and participant concerns (students, standardized patients, raters).

Results

Descriptive Statistics

Descriptive statistics (means), by case, are presented in *Table 2*. Based on the data gathering scores, all cases were of average difficulty, with case G being the most difficult. Compared with the history taking items, the physical examination items tended to be the most difficult. Averaged over the 11 stations that included a physical examination element, less than half of the items were attained. Global patient assessment mean scores, by case, ranged from 4.8 to 6.6. This variability may be due to differences in the communication challenge associated with each case or differences in the usage of the rating instrument by each standardized patient performing the case. The OMT scores were high, suggesting that the task was not too difficult and that the students were proficient. While there was some variability in the mean SOAP note scores, by case, the SOAP mean scores (averaged over the S, O, A, and P elements) were comparable with the SOAP global scores. This was not surprising in that for each note that was evaluated, both the global rating and the individual element scores (S, O, A, and P) were provided by the same rater.

Descriptive statistics summarized by student (ie, averaged over 12 encounters) are presented in *Table 3*. As mentioned previously, students were credited for less than half of the physical examination items. This may have been due to the difficulty of the task or the rigor with which the scoring criteria were applied (eg, maneuvers done correctly but not on skin—if required—were not credited). Nevertheless, averaged over the 11 encounters with a physical examination element, the maximum physical examination score was almost 80%, indicating that superior performance was possible. The variability in student scores for all scored components, with the exception of OMT, was not that great. Here, regardless of standardized patient prompts in the encounter and the presence of an osteopathic physician examiner in the room, some students either chose not to perform OMT or ran out of time.

Descriptive statistics by standardized patient and case are presented in *Table 4*. Here, only data for the COMLEX–USA–PE elements that are documented (data gathering) or evaluated (global patient assessment) by the standardized patient are provided. For most cases, the mean scores provided by the two standardized patients were reasonably close. This would be expected provided that there were not any meaningful ability differences between cohorts who were assessed by one standardized patient versus the other.* For some cases (eg, G), however, there was a large difference in mean data gathering scores between the two standardized patients performing the case. These differences may be due in part to random errors, training issues, checklist interpretation, standardized patient characteristics (eg, body mass), or standardized patient portrayal (eg, affect).

Case Performance

Quantitative summaries of case performance by COMLEX–USA–PE component are presented in *Table 5*. The case-total correlations indicate how well the encounter scores are able to discriminate between low- and high-ability students. Ideally, these values should be high, indicating that candidate performance on a specific case (eg, chest pain) is related to performance on the other cases (eg, headache, back

*The students were not assigned to test dates in any systematic way. Likewise, the standardized patients were rotated across the 11 assessment sessions. Therefore, one would not expect that students testing in one session would have higher or lower ability than students testing in another session.

Table 2
Descriptive Statistics by Case

	COMLEX–USA–PE Component									
Case	Medical History Taking*	Physical Examination*	PE-DO*	Data Gathering*	Osteopathic Manipulative Treatment†	Osteopathic Clinical Skills*	SOAP (mean)‡	SOAP (global)‡	Biomedical/ Biomechanical domain*	Global Patient Assessment‡
A	77.8	46.5	——	63.2	——	63.2	4.9	4.7	58.2	6.2
B	64.8	44.5	39.7	55.7	26.4	72.0	4.7	4.5	63.4	5.1
C	75.1	41.6	——	63.9	——	63.9	5.4	5.1	61.0	5.2
D	74.3	52.7	——	66.5	——	66.5	5.6	5.7	63.4	5.6
E	65.0	29.8	——	62.7	——	62.7	5.6	5.9	60.8	5.7
F	64.4	——	——	64.4	——	64.4	5.8	6.1	62.9	6.6
G	55.4	26.4	45.4	46.1	25.7	65.9	5.4	5.2	62.1	5.7
H	58.0	64.9	——	58.9	——	58.9	5.6	5.3	58.3	5.3
I	68.0	48.9	40.0	58.4	21.0	64.3	5.9	5.5	63.4	5.1
J	84.1	35.5	——	69.8	——	69.8	6.1	6.0	68.0	6.2
K	65.5	66.6	——	66.0	——	66.0	5.4	5.6	62.3	4.8
L	77.1	60.7	——	73.1	——	73.1	5.5	5.5	67.4	5.5
Total	**69.1**	**47.1**	**41.7**	**62.4**	**24.4**	**65.9**	**5.5**	**5.4**	**62.6**	**5.6**

*Percent score.
†Score 0-30.
‡Score 1-9.
PE-DO indicates physical examination checklist completed by osteopathic physicians for research purposes only; SOAP, subjective, objective, assessment, plan.

pain). If this were not the case, suggesting that student performance is strongly linked to the content of the case as opposed to the "generic" skills that are being measured, the reliability of the assessment would suffer. In addition, from an operational standpoint, it would be difficult to maintain the equivalence and comparability of multiple test forms needed for continuous testing.

The data presented in *Table 5* suggest that, at the case level, COMLEX–USA–PE produced skills-based scores that were able to discriminate between low and high performers. While discrimination indices were somewhat lower for the data gathering component, this was expected given the typical specificity of the case-based checklists that are used for scoring. The high discrimination values for the global patient assessment component indicate that interpersonal skills, at least for the types of cases used in this assessment, are not dependent on the specific patient complaint, ie, a student's ability in the humanistic domain (physician-patient communications and relationship) is reasonably case-invariant. The same conclusion can be drawn for the written communication component of COMLEX–USA–PE: The ability to summarize and interpret clinical data and formulate a differential diagnosis and plan is not overly dependent on the clinical scenario.

Correlations Among COMLEX–USA–PE Components

The correlations among COMLEX–USA–PE components are presented in *Table 6*. These correlations are based on aggregate student scores for each skill area. Osteopathic manipulative treatment scores were most highly correlated with biomedical/biomechanical ($r = 0.47$) and global patient assessment ($r = 0.46$) scores (OMT scores are part of the biomedical/biomechanical score, disattenuating this association). The moderate correlation between OMT and global patient assessment was expected given that a number of the measured traits (eg, respectfulness, professionalism) overlap. The correlation between global patient assessment and biomedical/biome-

Table 3
Descriptive Statistics (Student Level, n = 121)

Component	Mean	SD	Minimum	Maximum
History taking	69.1	5.8	51.9	81.1
Physical examination	47.1	11.6	23.5	79.6
PE-DO	41.7	16.9	10.0	79.3
Data gathering	62.4	6.0	45.4	77.8
Osteopathic manipulative treatment	24.4	4.9	1.7	30.0
Osteopathic clinical skills	65.9	6.0	49.0	82.0
SOAP (mean)	5.5	0.5	4.2	6.7
Subjective	5.5	0.6	3.5	6.8
Objective	5.6	0.7	3.3	7.1
Assessment	5.4	0.6	4.2	7.0
Plan	5.4	0.6	3.8	7.1
Biomedical	62.6	5.0	48.9	75.1
Global patient assessment (humanistic)	5.6	0.5	4.0	6.7

PE-DO indicates physical examination checklist completed by osteopathic physicians for research purposes only; SOAP, subjective, objective, assessment, plan.

chanical was moderate, with only 31% shared variance between measures.

Correlation of COMLEX–USA–PE Scores With Other Measures

Medical College Admission Test (MCAT), COMLEX–USA level 1 (initial attempt), and medical school grade point averages were available for 109 COMP students who participated in PSA. The grade point averages represent the cumulative grade point averages of all first- and second-year undergraduate medical coursework as well as the clinical rotation grades from the third year. In terms of MCAT scores, the only significant associations were between verbal reasoning and the COMLEX–USA–PE summary biomedical/biomechanical and global patient assessment scores (*Table 7*). Physical and biological science scores were not related to any COMLEX–USA–PE component or composite skills scores. This result was not unexpected given the time lag between assessments, the homogeneity of the PSA student sample, and the limited overlap in measurement domains.

The associations between COMLEX–USA level 1 and COMLEX–USA–PE scores were positive but small. The COMLEX–USA level 1 scores were not significantly related to OMT or global patient assessments. There was, however, a 9% shared variance in scores between level 1 and both the biomedical/biomechanical composite and the SOAP note scores. There were significant positive correlations between medical school grade point averages and all COMLEX–USA–PE component scores, with the exception of OMT.

The writing sample part of the MCAT is designed to assess, among other things, the ability of a candidate to synthesize concepts and ideas and to present ideas cohesively and logically. It is scored on an alphabetic scale ranging from J (lowest) to T (highest). The SOAP note exercise is also designed to assess a student's written communication skills, including the ability to document and synthesize patient data. Mean SOAP note scores stratified by MCAT writing sample scores are presented in *Table 8*. In general, the higher the MCAT writing sample score, the higher the mean SOAP note score.

Osteopathic Manipulative Treatment

Osteopathic manipulative treatment was assessed in 3 of the 12 encounters. Mean student performance for cases B, G, and I, on a percent-score metric, was 88.3 (SD, 14.0), 85.6 (SD, 18.7), and 70.2 (SD, 36.1), respectively. Overall, students did well on this part of the assessment. However, for cases B and I, the minimum student score was 0, which suggests that some students did not do any OMT. In 18 of 363 encounters, the student did not get credit on any item of the OMT scale (0 for all items). At the encounter level, OMT scores were most highly correlated with the global professional assessment scores provided by the standardized patients ($r = 0.29$). Among COMLEX–USA–PE components, the lowest correlation was between the physical examination checklist score and OMT ($r = -0.07$).

Table 4
Descriptive Statistics (Standardized Patients by Case)

Case	Standardized Patient	N	Data Gathering Mean (SD)	Global Patient Assessment Mean (SD)
A	17	68	62.0 (15.6)	6.4 (1.1)
	18	53	64.7 (11.5)	5.9 (0.7)
B	7	57	57.6 (12.7)	5.2 (0.7)
	8	64	53.9 (12.9)	5.0 (1.2)
C	1	65	66.8 (12.5)	5.4 (0.9)
	2	56	60.6 (15.9)	5.0 (0.8)
D	5	77	65.9 (14.3)	5.7 (0.6)
	6	44	67.5 (12.5)	5.5 (0.5)
E	9	64	62.8 (16.3)	5.6 (0.9)
	10	57	62.5 (14.7)	5.8 (1.0)
F	3	54	66.5 (18.2)	6.2 (1.2)
	4	67	62.7 (15.3)	6.9 (1.4)
G	13	67	50.8 (13.0)	5.9 (0.9)
	14	54	40.4 (13.4)	5.5 (0.6)
H	11	76	60.3 (9.9)	5.2 (0.6)
	12	45	56.5 (10.8)	5.5 (1.0)
I	23	11	66.4 (10.0)	6.0 (0.9)
	24	110	57.7 (14.4)	5.0 (0.8)
J	15	54	66.6 (11.7)	6.4 (0.8)
	16	67	72.4 (11.5)	6.1 (1.4)
K	19	56	65.0 (12.5)	5.0 (0.6)
	20	65	66.9 (11.3)	4.6 (0.8)
L	21	54	68.6 (9.0)	5.9 (0.7)
	22	67	76.8 (9.6)	5.3 (1.1)

Station Timing

The average time candidates spent in each examination station was tracked for select encounters. This was done to determine whether the 13-minute encounter time frame was adequate. A minimum of 24 encounters per case were timed. The average amount of time students spent in each examination station ranged from 7.5 minutes (case F) to 12.5 minutes (case B), well within the 13-minute time frame.

Reproducibility of COMLEX–USA–PE Encounter Scores

Generalizability coefficients were calculated for the COMLEX–USA–PE summary measures (biomedical/biomechanical, global patient assessment). The reliability of the biomedical/biomechanical composite over 12 encounters was 0.69 (standard error of measurement, 3.0). The reliability of the global patient assessment score (humanistic domain) over 12 encounters was 0.83 (standard error of measurement, 0.23). These coefficients, which do not account for case difficulty, choice of performing standardized patient, or potential variations in rater stringency, are comparable to those found for similar performance-based assessments (eg, Educational Commission for Foreign Medical Graduates/Clinical Skills Assessment; National Board of Medical Examiners standardized patient examination). Although the reproducibility of student biomedical/biomechanical scores (over 12 stations) was relatively low, especially compared with typical values for traditional high-stakes multiple-choice examinations, the reliability of this composite could be increased. Minimizing case specificity, lengthening the assessment, and weighting the component scores (eg, data gathering, OMT, SOAP) differently could all

Table 5
Case Discriminations (Case-Total Correlations)

Case	Data Gathering (History Taking and Physical Examination)	SOAP	Biomedical/ Biomechanical	Global Patient Assessment
A	0.16	0.50	0.42	0.60
B	0.31	0.54	0.50	0.57
C	0.43	0.45	0.54	0.43
D	0.30	0.46	0.41	0.34
E	0.23	0.44	0.46	0.38
F	0.20	0.59	0.40	0.42
G	0.28	0.44	0.53	0.46
H	0.22	0.63	0.43	0.43
I	0.30	0.48	0.51	0.55
J	0.36	0.54	0.60	0.55
K	0.30	0.45	0.46	0.36
L	0.37	0.45	0.52	0.60

SOAP indicates subjective, objective, assessment, plan.

Table 6
Component Correlations (Student Level, n = 121)

	History Taking	Physical Examination	Data Gathering	Osteopathic Manipulative Treatment	Osteopathic Clinical Skills	SOAP	Biomedical/ Biomechanical	Global Patient Assessment
History taking	1.00	0.33	0.86	0.19	0.80	0.33	0.77	0.53
Physical examination	——	1.00	0.71	0.10	0.65	0.24	0.62	0.20
Data gathering	——	——	1.00	0.19	0.94	0.30	0.86	0.47
Osteopathic manipulative treatment	——	——	——	1.00	0.50	0.18	0.47	0.46
Osteopathic clinical skills	——	——	——	——	1.00	0.32	0.92	0.56
SOAP	——	——	——	——	——	1.00	0.66	0.30
Biomedical/ biomechanical	——	——	——	——	——	——	1.00	0.56
Global patient assessment	——	——	——	——	——	——	——	1.00

SOAP indicates subjective, objective, assessment, plan.

Table 7
Correlation of COMLEX–USA–PE Component Scores With Other Measures

	Data Gathering	Osteopathic Manipulative Treatment	Osteopathic Clinical Skills	SOAP	Biomedical/ Biomechanical	Global Patient Assessment
Medical College Admission Test						
Verbal reasoning	0.13	0.16	0.18	0.12	0.19*	0.20*
Physical sciences	−0.07	0.09	−0.03	0.03	−0.01	−0.04
Biological sciences	−0.15	0.10	−0.10	0.10	−0.04	−0.05
COMLEX–USA						
Level 1	0.21*	0.12	0.22*	0.32*	0.30*	0.09
Medical school grade point average	0.29*	0.12	0.29*	0.36*	0.38*	0.11*

*$P < .05$.
COMLEX-USA indicates Comprehensive Osteopathic Medical Licensing Examination–USA; SOAP, subjective, objective, assessment, plan.

serve to decrease measurement error. The global patient assessment scores were less prone to measurement errors, indicating a consistency in student interpersonal performances over patient encounters.

Quality Assurance

Quality assurance was an integral part of PSA. For data gathering (history taking, clinical skills examination), approximately one third (n = 474 observations) of the encounters were viewed and scored by a second standardized patient. Overall, based on a mean checklist length of 18.9 items, there was an average of 2.5 (minimum, 0; maximum, 8) discrepancies per encounter.

Postexamination Survey

Candidates, osteopathic physician examiners, and standardized patients all reported that on average, the 13-minute encounter time limit was adequate. Students indicated that they wanted more time to complete the SOAP notes. However, osteopathic physicians rating the SOAP notes reported that only 3 of more than 1450 notes had sections that were incomplete, suggesting that time pressure for the written exercise was minimal.

The students overwhelmingly reported that the cases were realistic and appropriately challenging. Students, standardized patients, and osteopathic physician examiners admitted to having some fatigue, most notably at the end of the daily session of 12 encounters. The standardized patients did not report any adverse effects from the physical examination maneuvers or OMT performed by the students. Few checklist rating problems were reported by standardized patients or osteopathic physician examiners, except for numerous reports by osteopathic physician examiners citing difficulties in scoring OMT while remaining seated in the examination rooms.

Discussion

The use of objective structured clinical examinations and standardized patient assessments has increased dramatically since the initial work of Harden and Gleeson.[20] Performance-based assessments are now commonly used by licensing and board-certification bodies to assess the professional competencies of various examinee groups. The Educational Commission for Foreign Medical Graduates administers the Clinical Skills Assessment as part of the certification requirements for graduates of international medical schools. Certification, which includes the assessment of clinical skills, is meant to ensure that graduates of international medical schools (ie, outside of the United States and Canada) are ready to enter graduate medical education programs in the United States. The Medical Council of Canada evaluates clinical skills as part of the licensure requirements for physicians wishing to practice in Canada. Finally, the National Board of Medical Examiners has administered standardized patient cases to thousands of examinees over the past decade in preparation for inclusion of a clinical skills component into the United States Medical Licensing Examination.[21] While the psychometric adequacy of these types of assessments has been studied extensively, validation is an ongoing process, requiring regular accumulation of evidence to support the use of test scores. For newly developed assessments such as COMLEX–USA–PE, the data gathered from pilot tests can aid in the validation process.

The reliability of the student scores over the 12-station assessment was modest, but in accordance with what has been reported in other studies.[6,8,22,23] Although no performance standards have been set for COMLEX–USA–PE, it would appear that the biomedical/biomechanical and humanistic composite scores, while somewhat prone to task sampling and rater (or standardized patient) error sources, can be used

Table 8
SOAP Note Scores Stratified by Medical College Admission Test Writing Sample Levels

Medical College Admission Test level	SOAP Note Score			
	N	Mean (SD)	Minimum	Maximum
L	4	4.9 (0.5)	4.2	5.3
M	22	5.3 (0.5)	4.4	6.1
N	10	5.5 (0.5)	4.9	6.3
O	13	5.5 (0.4)	4.7	6.0
P	19	5.3 (0.5)	4.5	6.4
Q	22	5.5 (0.4)	4.8	6.5
R	10	5.8 (0.4)	5.2	6.4
S	8	6.0 (0.5)	5.5	6.7
T	1	*	*	*

*Insufficient n to report.
SOAP indicates subjective, objective, assessment, plan.

to make reliable decisions regarding an osteopathic medical student's readiness to enter graduate medical education. Choosing appropriate encounters (ie, cases in which content specificity is minimized), enhancing standardized patient/rater training, and providing "real-time" quality assurance data and feedback could further enhance the generalizability of COMLEX–USA–PE scores.

Unexpectedly, we found that for some cases, the choice of standardized patient had a measurable impact on average student scores. Nevertheless, as long as the between-case and between–standardized patient differences are more or less constant, it is possible not only to adjust student scores to take into account the relative difficulty of the case, but also the relative stringency of the particular standardized patient who performed it.[10] In addition, where "real-time" quality assurance data are available, training and checklist interpretation problems can be rectified immediately. This will mitigate any variability in student scores that could be attributed to the choice of standardized patient. Finally, the adoption of alternate scoring models and score-equating strategies could aid in minimizing the potential impact of selecting different cases and/or raters (osteopathic physicians, standardized patients) for various test forms.

The availability of quality assurance observations and secondary data is important for score validation. Similar to results reported in previous studies,[24,25] we found that there was some error in the documentation of history taking and physical examination. Overall, there were approximately 2.5 disagreements on the average 19-item checklist. Nevertheless, the discrepancy rate represents disagreement in either direction (ie, standardized patient giving credit, observer not giving credit; not giving credit, observer giving credit). It is important to note that the ability to document student questioning and actions via monitor can be impeded by poor acoustics (eg, soft-spoken student) or less-than-ideal camera angles (eg, student blocking camera view). In addition, if discrepancies occur at random, the overall impact of documentation errors will be minimized over 12 encounters. While scoring and interpretation errors are inevitable, it is important to understand what caused them. Here, the availability of a large bank of case videos will be useful, as these can be used to study the fidelity of case portrayals and the accuracy of scoring (both for standardized patients and osteopathic physicians). This video library can also be used to gather training material (eg, benchmark performances) and provide performance samples for standard-setting exercises. Based on our results, additional studies focusing on determining the nature of documentation errors are warranted.

Although there is marked variability in the time practicing physicians spend soliciting information from patients and performing physical examinations, the logistics of examination administration make it difficult to vary station timing. Therefore, it is essential that candidates have sufficient time to complete the necessary tasks. The purpose of COMLEX–USA–PE is not to assess the ability to perform the tasks rapidly. If the patient encounter time were not sufficient, then decisions regarding examinees' readiness would be confounded by their ability to perform the necessary tasks quickly.[26] This would result in a speeded performance assessment, potentially compromising the validity and reliability of test scores. Based on our data, the 13-minute encounter time frame for COMLEX–USA–PE appears on average to be adequate. As a result, the

examinee's working rate will not systematically influence his or her performance. Nevertheless, given the complexity of patient interviews and assessments, a more focused exploration of timing issues, especially in relation to case content and candidate factors, would be informative.

Students performed reasonably well on PSA with the possible exception of physical examination skills. For physical examination skills, students were required not only to perform the maneuver, but also to do it correctly. Physical examination maneuvers completed but not done correctly (eg, on the skin) were not credited. The presence of an osteopathic physician examiner in the room may have prompted the students to perform OMT, possibly resulting in high-average OMT scores. However, in these stations, OMT may have been done at the expense of physical examination maneuvers. Nevertheless, some students chose not to do OMT because of skill deficits or lack of time. Given the fundamental role of OMT in osteopathic medical training, the reasons for not treating the patient will need to be explored in future studies. Taking the osteopathic physician examiner out of the room will add to the verisimilitude of the assessment and significantly decrease the logistic complexity of examination administration as well as provide for more valid and reproducible scores. Overall, the case performance data suggested that COMLEX–USA–PE is able to provide scores that are able to discriminate between low- and high-ability students. Although case-specificity plays a role in standardized patient–based performance assessments,[27] the careful development and screening of appropriate performance scenarios will result in an assessment that has an adequate representation of patient complaints and also yields scores with desirable measurement properties.

Analyses of the internal structure of COMLEX–USA–PE scores provide additional evidence that the test components conform to the construct on which the proposed assessment interpretations are based. The biomedical/biomechanical and humanistic domains were only moderately related. This indicates that, as expected, a physician's skill in relating to patients may differ from his or her ability to recognize disease patterns, generate differential diagnoses, synthesize medical data, or treat specific conditions. This also suggests that if a total score (ie, biomedical/biomechanical and global patient assessment) were used for assessment decisions, it would be likely that some candidates could compensate for substandard biomedical/biomechanical skills with superior humanistic qualities and vice versa.

Based on the prevailing literature,[28] the limited strength of the associations between COMLEX–USA–PE component/composite scores and MCAT, COMLEX-USA level 1, or medical school aptitude and ability measures was expected. Although some basic medical science proficiency (eg, understanding the mechanisms of medical problems and disease processes) is required for patient workup and management, COMLEX–USA–PE was designed to measure what a candidate can do, not necessarily what he or she knows. The low correlations with external measures, which can be attributed to the somewhat independent nature of the constructs being measured and the time lag between assessments, provide evidence for the discriminant validity of the assessment scores.

Gathering validity evidence for COMLEX–USA–PE is an ongoing process. Additional measures need to be collected (eg, COMLEX–USA levels 2 and 3 scores, residency evaluations) and checked for their association with COMLEX–USA–PE scores. Ultimately, it would be expected that more able students, as delimited by COMLEX–USA–PE, would perform at higher levels in residency programs. In addition, issues related to differential case functioning have yet to be adequately explored. Analysis of PSA data by various standardized patient and student characteristics (eg, gender, age) will provide additional evidence to support the fairness of the scores.

Conclusion

The results from PSA provide valuable data that can be used to guide the possible implementation of a clinical skills assessment for graduates of colleges of osteopathic medicine. The analysis of data from 1452 standardized patient encounters provides additional evidence to support the use of COMLEX–USA–PE for assessing the readiness of osteopathic medical students to provide patient care in supervised graduate medical education training programs. Although the reproducibility of COMLEX–USA–PE scores was less than that typically achieved for high-stakes board-certification and licensure examinations, it is comparable to that found for other standardized patient assessments.

References

1. Whelan GP. Educational Commission for Foreign Medical Graduates: Clinical Skills Assessment prototype. *Med Teach*. 1999;21(2):156-160.

2. Ben David MF, Klass DJ, Boulet J, De Champlain A, King AM, Pohl HS, et al. The performance of foreign medical graduates on the National Board of Medical Examiners (NBME) standardized patient examination prototype: A collaborative study of the NBME and the Educational Commission for Foreign Medical Graduates (ECFMG). *Med Educ*. 1999;33(6):439-446.

3. Medical Council of Canada. Qualifying Examination Part II, Information Pamphlet. Ottawa, Ontario, Canada: Medical Council of Canada; 2002.

4. Anderson MB, Stillman PL, Wang Y. Growing use of standardized patients in teaching and evaluation in medical education. *Teach Learn Med*. 1994;6(1):15-22.

5. Klass DJ. "High-stakes" testing of medical students using standardized patients. *Teach Learn Med*. 1994;6(1):28-32.

6. Boulet JR, Ben David MF, Ziv A, Burdick WP, Curtis M, Peitzman S, et al. Using standardized patients to assess the interpersonal skills of physicians. *Acad Med*. 1998;73(10 Suppl):S94-S96.

7. Gomez JM, Prieto L, Pujol R, Arbizu T, Vilar L, Pi F, et al. Clinical skills assessment with standardized patients. *Med Educ*. 1997;31(2):94-98.

8. Vu NV, Barrows HS. Use of standardized patients in clinical assessments: Recent developments and measurement findings. *Educ Res*. 1994;23(3):23-30.

9. van der Vleuten C, Swanson DB. Assessment of clinical skills with standardized patients: State of the art. *Teach Learn Med*. 1990;2(2):58-76.

10. Swanson DB, Clauser BE, Case SM. Clinical skills assessment with standardized patients in high-stakes tests: A framework for thinking about score precision, equating, and security. *Adv Health Sci Educ Theory Pract*. 1999;4:67-106.

11. Reznick RK, Blackmore D, Dauphinee WD, Rothman AI, Smee S. Large-scale high-stakes testing with an OSCE: Report from the Medical Council of Canada. *Acad Med*. 1996;71(1 Suppl):S19-S21.

12. Whelan GP. Educational Commission for Foreign Medical Graduates: Lessons learned in a high-stakes, high-volume medical performance examination. *Med Teach*. 2000;22(3):293-296.

13. Heun L, Brandau DT, Chi X, Wang P, Kangas J. Validation of computer-mediated open-ended standardized patient assessments. *Int J Med Inf*. 1998; 50(1-3):235-241.

14. Portanova R, Adelman M, Jollick JD, Schuler S, Modrzakowski M, Soper E, et al. Student assessment in the Ohio University College of Osteopathic Medicine CORE system: Progress testing and objective structured clinical examinations. *J Am Osteopath Assoc*. 2000;100(11):707-712.

15. Heun L, Brandau DT, Chi X, Wang P, Kangas J. Validation of computer-mediated open-ended standardized patient assessments. *Int J Med Inf*. 1998; 50(1-3):235-241.

16. Shen L, Cavalieri T, Clearfield M, Smoley J. Comparing medical knowledge of osteopathic medical trainees in DO and MD programs: A random effect meta-analysis. *J Am Osteopath Assoc*. 1997;97(6):359-362.

17. Graneto J. Testing osteopathic medical school graduates for licensure: Is COMLEX–USA the most appropriate examination? *J Am Osteopath Assoc*. 2001;101(1):26-32.

18. Meoli FG, Cavalieri T, Buser B, Smoley J, Shen L. National Board of Osteopathic Medical Examiners in the 21st century. *J Am Osteopath Assoc*. 2000; 100(11):703-706.

19.. Brennan RL. Elements of generalizability theory. Iowa City, Iowa: ACT Publications; 1992.

20. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ*. 1979;13(1):41-54.

21. Hallock JA. ECFMG and the Challenges Facing International Medical Graduates. Washington, DC: Association of American Medical Colleges; 2002.

22. Cohen DS, Colliver JA, Robbs RS, Swartz MH. A large-scale study of the reliabilities of checklist scores and ratings of interpersonal and communication skills evaluated on a standardized-patient examination. *Adv Health Sci Educ*. 1997;1:209-213.

23. Boulet JR, Friedman Ben-David M, Ziv A, Burdick WP, Gary NE. The use of holistic scoring for post-encounter written exercises. Melnick D, ed. *Proceedings of the Eighth Ottawa Conference of Medical Education and Assessment*. Philadelphia, Pa: National Board of Medical Examiners; 2000.

24. De Champlain AF, Margolis MJ, King A, Klass DJ. Standardized patients' accuracy in recording examinees' behaviors using checklists. *Acad Med*. 1997;72(10 Suppl 1):S85-S87.

25. De Champlain AF, Macmillan MK, Margolis MJ, King AM, Klass DJ. Do discrepancies in standardized patients' checklist recording affect case and examination mastery-level decisions? *Acad Med* 1998;73(10 Suppl):S75-S77.

26. Chambers KA, Boulet JR, Gary NE. The management of patient encounter time in a high-stakes assessment using standardized patients. *Med Educ*. 2000;34(10):813-817.

27. van der Vleuten CP, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: Issues of reliability. *Med Educ*. 1991;25(2):110-118.

28. Ayers WR, Boulet JR. Establishing the validity of test score inferences: Performance of 4th-year US medical students on the ECFMG Clinical Skills Assessment. *Teach Learn Med*. 2001;13(4):214-220.

OFFICIAL CALL

To the Officers and Members of the American Osteopathic Association:

You are hereby notified that the annual business meetings of the American Osteopathic Association will be held July 15-20, 2003, at the Fairmont Hotel in Chicago.

The annual meeting of the Board of Trustees will begin at 8:30 AM on Tuesday, July 15.

The House of Delegates will convene for the annual business session of the association at 9:30 AM on Friday, July 18. The House's Committee on Credentials will register delegates and alternate delegates from 2 PM to 6 PM on Thursday, July 17, and from 7 AM to 9 AM on Friday, July 18. The House will conclude its session on Sunday, July 20.

The executive director of each state osteopathic medical association, the executive director of each osteopathic specialty college, the chairman of the Executive Committee of the AOA Bureau of Interns and Residents, and the administrator of the Student Osteopathic Medical Association shall certify the names of their delegates and alternate delegates to the AOA's executive director at least 30 days prior to the House of Delegates' meeting.

Anthony A. Minissale, DO, President
Mark A. Baker, DO, Speaker of the House