Relationships Between Scores on the COMLEX-USA Level 2-Performance Evaluation and Selected School-Based Performance Measures

Helen H. Baker, PhD, MBA; Michael K. Cope, PhD; Michael D. Adelman, DO, JD; Stephanie Schuler; Robert W. Foster, DO; and John R. Gimpel, DO, MEd

The Comprehensive Osteopathic Medical Licensing Examination USA Level 2 Performance Evaluation (COMLEX-USA Level 2-PE) is a national multistation performance examination designed to examine students’ osteopathic clinical skills. The current study examines the relationship between achievement levels on the COMLEX-USA Level 2-PE and selected school-related variables for the class of 2005 at the West Virginia School of Osteopathic Medicine in Lewisburg, WV (N=70). Significant (P<.01) correlations between the COMLEX-USA Level 2-PE summary performance and selected academic achievement measures include: weighted Physical Diagnosis grade, 0.41; weighted year 1 and year 2 Osteopathic Principles and Practice grade, 0.37; overall year 2 grade point average, 0.42; the objective structured clinical evaluation (OSCE) Physical Examination score, 0.40; and the OSCE Total Station score, 0.33. While further research is needed, the current study found modest but notable relationships between school-generated academic variables and performance on the COMLEX-USA Level 2-PE, and therefore supports the validity of the COMLEX-USA Level 2-PE examination for assessing the clinical skills of future osteopathic physicians.

http://www.joa.org/content/vol109/issue5/

Valid and reliable assessment of medical student performance has long been a major challenge.1,2 Recent focus on medical errors,3 and evidence that improvement in the quality of medical care requires reform in the ways physicians are educated and evaluated,4 have led to the conclusion that medical schools, accrediting bodies, and licensing boards all have responsibilities in ensuring the competence of medical school graduates.4

A full explanation of validity, reliability, and test validation requires further reading.5,6 However, to understand the present study, a brief review of terminology used in educational evaluation may be helpful.

The reliability of a test refers to the consistency of results. While there are several different ways test reliability can be assessed, in this article, the term reliability refers to the degree of internal consistency of the measurement. When educators assign students to a 12-station examination, it is expected that high-performing students will consistently perform well on most of these stations, and low-performing students will consistently demonstrate poor performance on most stations. The degree to which particular students consistently perform well (or poorly) across all stations is reflected in the associated reliability statistic. Reliability is necessary but not sufficient for test validity.

Validity refers to how appropriate the interpretation of results is. If an examination is intended to provide information to osteopathic medical educators and the general public regarding whether a student has the skills and abilities necessary for osteopathic clinical practice, then the test’s validity refers to the degree to which conclusions based on this examination accurately reflect the underlying concept of “the skills and abilities necessary for osteopathic clinical practice.” Some attitudes, skills, and abilities can be assessed in a 1- or 2-day examination, and others (for example, lifelong learning skills, and work habits when one believes one is not being observed) probably cannot. Thus, test validity is always a matter of degree.

A statistic useful in determining test validity is a correlation. A correlation coefficient expresses the strength of the relationship between two sets of scores, and ranges from +1.00 (a perfect positive correlation) to zero (no correlation) to −1.00 (a perfect negative relationship). The square of a correlation is the proportion of variation in one variable accounted for by the other: for example, if the correlation between X and Y is 0.40, then (calculating the square, 0.40 × 0.40 = 0.16) 16% of the variation in Y can be accounted for by knowing X. While most biomedical researchers would consider a correlation of 0.90 to be strong and a correlation of 0.05 to be weak, different disciplines (eg, physiology, psychology) have developed different traditions regarding how high a correlation must be to have practical significance.7 In this article, correlations above 0.35 are considered moderate and...
of some practical importance, though they account for a modest portion of the variation. Furthermore, if a test has low reliability, correlations between that test and any other variable will also be low. A test's reliability coefficient is often seen as the upper limit of how well a test could possibly correlate with anything else.

In 1959, Campbell and Fiske\(^8\) wrote a classic article regarding test validity, in which they (in part) outlined the idea of convergent and discriminant validity. In the current context, for example, there might be two measures of osteopathic clinical performance: a school-constructed clinical performance examination and a national licensing examination. According to the tenets of Campbell and Fiske\(^8\) if two examinations are intended to assess the same skills and abilities, a positive correlation should be found between these two measures. To the extent that the school examination focuses on school-specific outcomes (for example, emphasizes diseases common in Appalachia) while the national examination focuses on other problems (for example, diseases in the Hispanic population, a group underrepresented in Appalachia), this correlation would most likely be less than perfect. A lower or zero correlation would be expected between a national clinical performance examination and a basic science multiple-choice test.

At least since the work of Barrows and Abrahamson\(^9\) in 1964, medical educators have been using some form of standardized patients as part of the student assessment process.\(^3,10\) While multiple-choice tests continue to be valuable in assessing students' medical knowledge and clinical reasoning, the use of standardized patients allows for a more direct assessment of clinical performance and student abilities to interact effectively with patients. Numerous validity studies are reported in the literature regarding high-stakes clinical skills examinations, including the Educational Commission for Foreign Medical Graduates Clinical Skills Assessment\(^11,12\) and clinical skills assessments by the Medical Council of Canada.\(^13\)

Two national clinical skills assessments for osteopathic physicians in training were implemented in 2004: the Comprehensive Osteopathic Medical Licensing Examination USA Level 2 Performance Evaluation (COMLEX-USA Level 2-PE) by the National Board of Osteopathic Medical Examiners (NBOME),\(^14\) and the United States Medical Licensing Examination Step 2 Clinical Skills by the National Board of Medical Examiners.\(^15\) All physicians graduating from US medical schools in 2005 or thereafter, as well as international medical school graduates, are now required to pass a national clinical skills assessment for licensure in the United States.

Specific to osteopathic medical education, the NBOME has developed and implemented the COMLEX-USA Level 2-PE, a national multistation examination designed to evaluate trainees' osteopathic clinical skills.\(^16\) Passing this examination is a requirement for licensure through the COMLEX-USA pathway. Furthermore, according to the American Osteopathic Association's accreditation standards,\(^17\) all students [in colleges of osteopathic medicine] must take and pass the National Board of Osteopathic Medical Examiners, Inc. (NBOME) Comprehensive Osteopathic Medical Licensing Examination (COMLEX) Level I prior to graduation. All students must take COMLEX Level II Cognitive Evaluation (CE) and Performance Evaluation (PE) components prior to graduation. All students who enter in the 2004–2005 academic year, and all students who graduate after December 1, 2007, must also pass NBOME Cognitive Evaluation (CE) and Performance Evaluation (PE) components of COMLEX Level II prior to graduation.\(^17\)

The validity of this examination will have an impact on graduation and licensure of all future osteopathic physicians. The COMLEX-USA Level 2-PE is a 12-station, standardized patient–based clinical skills assessment administered at NBOME’s National Center for Clinical Skills Testing in Conshohocken, Pa. The examination assesses the clinical skills of osteopathic trainees in two distinct domains. The Humanistic domain evaluates physician-patient communication, interpersonal skills, and professionalism. The Biomedical/Biomechanical domain assesses data gathering (history taking and physical examination skills); osteopathic principles and osteopathic manipulative treatment; and written communication skills, including synthesis of clinical findings, integrated differential diagnosis, and formulation of a diagnostic and treatment plan. These patient-centered skills are evaluated in the context of clinical encounters with standardized patients.\(^16\)

While a candidate can make up for substandard performance across individual stations or component skills within each domain, a candidate cannot compensate across domains. Therefore, candidates must receive a passing score in both domains to receive a passing score for the COMLEX-USA Level 2-PE. For example, a candidate who fails to meet defined performance standards in physician-patient communication, interpersonal skills, and professionalism would receive a failing score for the COMLEX-USA Level 2-PE despite meeting the standards for history taking and PE, and vice versa.

The results of a series of studies on the new COMLEX-USA Level 2-PE have been encouraging in terms of reliability and validity.\(^18,19\) However, because the examination was first administered in 2004–2005, no studies are available that relate actual performance on the new PE to performance in osteopathic medical school. The current study investigates the relationship between achievement on the COMLEX-USA Level 2-PE and selected school-related variables for the class of 2005 at the West Virginia School of Osteopathic Medicine (WVSOM) in Lewisburg. This group was the first graduating class at WVSOM to take this national licensing examination.
Methods
Approval was obtained from WVSOM’s institutional review board to conduct this research using existing educational data. Subjects for the study were 70 students (31 men, 39 women) who were in the class of 2005 (as composed in April 2004), and who took WVSOM’s third-year objective structured clinical evaluation (OSCE) and subsequently the COMLEX-USA Level 2-CE and Level 2-PE. One student who took the third-year OSCE did not subsequently take the COMLEX-USA Level 2-PE or graduate, for reasons unrelated to the student’s performance on the OSCE. An additional two students who completed both examinations and therefore are included in this analysis also did not graduate in May 2005, for reasons not related to performance on the COMLEX-USA Level 2-PE.

COMLEX-USA Level 2-PE
All subjects were required by the school to participate in, but not necessarily to pass, the COMLEX-USA Level 2-PE. For this examination, though candidates received additional performance feedback from the NBOME directly, schools were advised only regarding the passing or failing result for each student. Accordingly, these data were coded “1” for pass and “0” for fail.

COMLEX-USA Level 1 and Level 2-CE
All subjects were required to pass the COMLEX-USA Level 1 written test and Level 2-CE to graduate. In addition to pass-fail information, data reported by the NBOME for the schools regarding NBOME’s written tests included numeric standard scores (for example, a score value of 505, not just “pass”). These numeric values were used in the current analysis. When a student failed the examination on the first attempt and subsequently repeated it, the score on the first attempt was used in this analysis.

Grades
All subjects participated in WVSOM’s systems-based curriculum. (The first class to participate in the new problem-based learning program will graduate in 2007, and has not yet taken the COMLEX-USA Level 2-PE.) Most courses in the first 2 years of the WVSOM curriculum use a standard numerical grading scale. When a student failed a course and was required to repeat it, the first (failing) numeric score was used in the current analysis. The following variables were calculated from raw numeric course grades:

- year 1 overall grade point average (GPA);
- year 2 overall GPA;
- weighted total of WVSOM’s two Physical Diagnosis course grades; and
- weighted total from the first year and year 2 Osteopathic Principles and Practice course grades.

These derived variables were calculated using standard credit hour calculations—for example, for the four semesters of Osteopathic Principles and Practice, the course weightings are 2.25, 2.25, 2.5, and 1.5 hours, respectively—multiplied by the numeric student grade, then divided by the number of credit hours. The overall GPA is the weighted average GPA for the entire curriculum, as calculated by the registrar’s office.

Third-Year Objective Structured Clinical Evaluation
To prepare for this national licensing examination and to meet other goals identified by the administration and faculty, WVSOM developed a third-year OSCE, which was first implemented in the 2003–2004 academic year with the class of 2005. This 12-station examination was administered on 3 half-days in April 2004. At each of the 12 stations, students were given 13 minutes to take a patient history and perform a focused physical examination, and then had 9 minutes to complete the COMLEX-USA Level 2-PE or graduate, for reasons unrelated to the student’s total History and Physical score; the Physical Examination score; and the SOAP Note form score. In addition, the standardized patient and/or in-room evaluator provided global ratings of Communication and Professionalism.

Statistical Analysis
Because COMLEX-USA Level 2-PE scores were reported to the colleges of osteopathic medicine as a single pass or fail for each student, these data are dichotomous, while the other variables being studied are continuous. Point-biserial correlations are used to measure the strength of the relationship between a dichotomous variable and a continuous variable; accordingly, point-biserial correlations were calculated, using SPSS statistical software (13.0 for Windows; SPSS Inc, Chicago, Ill).

Results
Point-biserial correlations with selected academic performance measures are presented in Table. The correlation of COMLEX-USA Level 2-PE with the weighted total numeric grade in the Physical Diagnosis course sequence was 0.41, while correlation with performance in the Osteopathic Principles and Practice course sequence was 0.37; both of these correlations were statistically significant at the .01 level. Correlations with GPAs were: year 1, 0.28 ($P < .05$); year 2, 0.42 ($P < .01$); and years 3 and 4 (combined), 0.30 ($P < .05$). The inter-station reliability (calculated via Cronbach α) for each third-year OSCE measure ranged from 0.36 to 0.76. The correlation of the COMLEX-USA Level 2-PE with the Physical Examination score on the third-year OSCE was 0.40 ($P < .01$), and the
of getting strong correlations. Furthermore, because few students failed the examination, it was expected that correlations with COMLEX-USA Level 2-PE might be small. The reliability of the third-year OSCE was less than optimal, and reliability statistics are not available for course grades or overall GPA for any year of WVSOM’s curriculum. Given these statistical constraints, the correlations obtained are encouraging.

The Physical Diagnosis and Osteopathic Principles and Practice sequences were the two courses in years 1 and 2 of the curriculum that included substantial, faculty-evaluated clinical skills components. Moderate correlations (0.41 and 0.37) were found between the weighted grades in these curriculum components and the COMLEX-USA Level 2-PE administered to these students in year 4. Further analysis is needed regarding the specific grade components within such courses that most contribute to the relationship, and specifically whether multiple-choice test performance in these courses vs clinical skills, as evaluated by faculty, best predict subsequent performance on the COMLEX-USA Level 2-PE. The apparent higher correlation for year 2 (0.42, the highest correlation of any in this data set) may reflect the increased clinical content in the year 2 curriculum, compared with a primarily basic science orientation during the year 1 curriculum (0.28). While other academic factors, such as quality of oral and written case presentations and laboratory and clinical examinations, have some minimal influence on course grades at WVSOM, the class ranking for year 1 and year 2 GPA is primarily determined by performance on multiple choice tests, and partially by course grades in Physical Diagnosis and Osteopathic Principles and Practice. Therefore, the correlations obtained between GPAs and subsequent PE scores seem very appropriate.

The correlation between overall year 3 and 4 GPA and PE was 0.30. While some of WVSOM’s clinical rotations include multiple-choice tests and graded osteopathic manipulative medicine case reports, the clinical years’ GPA is primarily determined by global ratings of student performance by adjunct clinical faculty. The clinical rotations are highly diverse; for example, the skills and abilities needed for a high grade in an 8-week rural family medicine rotation may be very different from those needed during a 4-week pathology elective at a tertiary medical center. While further analysis is

<table>
<thead>
<tr>
<th>Variables</th>
<th>Cronbach α</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Diagnosis</td>
<td>...</td>
<td>0.41†</td>
</tr>
<tr>
<td>Osteopathic Principles and Practice</td>
<td>...</td>
<td>0.37†</td>
</tr>
<tr>
<td>GPA Year 1</td>
<td>...</td>
<td>0.28‡</td>
</tr>
<tr>
<td>GPA Year 2</td>
<td>...</td>
<td>0.42‡</td>
</tr>
<tr>
<td>GPA Years 3 and 4§</td>
<td>...</td>
<td>0.30‡</td>
</tr>
<tr>
<td>OSCE □</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical Examination</td>
<td>.69</td>
<td>0.40†</td>
</tr>
<tr>
<td>History</td>
<td>.71</td>
<td>0.23</td>
</tr>
<tr>
<td>SOAP Note form</td>
<td>.56</td>
<td>0.12</td>
</tr>
<tr>
<td>Total Station score</td>
<td>.76</td>
<td>0.33‡</td>
</tr>
<tr>
<td>Global Ratings □</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communication</td>
<td>.36</td>
<td>0.29‡</td>
</tr>
<tr>
<td>Professionalism</td>
<td>.63</td>
<td>0.24‡</td>
</tr>
<tr>
<td>COMLEX-USA CE □</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>...</td>
<td>0.08</td>
</tr>
<tr>
<td>Level 2</td>
<td>...</td>
<td>0.29‡</td>
</tr>
</tbody>
</table>

* CE indicates cognitive evaluation (written); COMLEX, Comprehensive Osteopathic Medical Licensing Examination; GPA, grade point average; PE, performance evaluation; SOAP, subjective, objective, assessment, and plan.
† P<.01.
‡ P<.05.
§ Years 3 and 4 are integrated and therefore reported together.
// Objective Structured Clinical Evaluation; this test is administered once near the end of the third clinical year.

Correlation with the Total Station score (a weighted average of the Physical Examination, History, and SOAP Note ratings) was 0.33 (P<.01). The correlation with the OSCE Communication score was 0.29 (P<.05), and the correlation with the Professionalism score was 0.24 (P<.05). The correlation of COMLEX-USA Level 2-PE with COMLEX-USA Level 2-CE was 0.29, (P<.05), while the correlation with COMLEX-USA Level 1 written test was not statistically significant.

Discussion
To pass the COMLEX-USA Level 2-PE, students must pass both the Humanistic domain and the Biomedical/Biomechanical domain. In 2005, the colleges of osteopathic medicine received only pass or fail data (ie, not numeric scores). Summarizing performance on two dimensions into a single, dichotomous score placed severe restraint on the possibility of getting strong correlations. Furthermore, because few students failed the examination, it was expected that correlations with COMLEX-USA Level 2-PE might be small. The reliability of the third-year OSCE was less than optimal, and reliability statistics are not available for course grades or overall GPA for any year of WVSOM’s curriculum. Given these statistical constraints, the correlations obtained are encouraging.

The Physical Diagnosis and Osteopathic Principles and Practice sequences were the two courses in years 1 and 2 of the curriculum that included substantial, faculty-evaluated clinical skills components. Moderate correlations (0.41 and 0.37) were found between the weighted grades in these curriculum components and the COMLEX-USA Level 2-PE administered to these students in year 4. Further analysis is needed regarding the specific grade components within such courses that most contribute to the relationship, and specifically whether multiple-choice test performance in these courses vs clinical skills, as evaluated by faculty, best predict subsequent performance on the COMLEX-USA Level 2-PE. The apparent higher correlation for year 2 (0.42, the highest correlation of any in this data set) may reflect the increased clinical content in the year 2 curriculum, compared with a primarily basic science orientation during the year 1 curriculum (0.28). While other academic factors, such as quality of oral and written case presentations and laboratory and clinical examinations, have some minimal influence on course grades at WVSOM, the class ranking for year 1 and year 2 GPA is primarily determined by performance on multiple choice tests, and partially by course grades in Physical Diagnosis and Osteopathic Principles and Practice. Therefore, the correlations obtained between GPAs and subsequent PE scores seem very appropriate.

The correlation between overall year 3 and 4 GPA and PE was 0.30. While some of WVSOM’s clinical rotations include multiple-choice tests and graded osteopathic manipulative medicine case reports, the clinical years’ GPA is primarily determined by global ratings of student performance by adjunct clinical faculty. The clinical rotations are highly diverse; for example, the skills and abilities needed for a high grade in an 8-week rural family medicine rotation may be very different from those needed during a 4-week pathology elective at a tertiary medical center. While further analysis is
needed, we believe that this relatively low correlation with COMLEX-USA Level 2-PE resulted from a lack of homogeneity in clinical grades, reflecting both true differences in expected performance on different rotations, and error variance in the assignment of global ratings. Colleges of osteopathic medicine that assign students primarily to a single tertiary medical center with a limited number of faculty, and/or that place heavy weight on multiple choice tests, might obtain substantially different results.

The low correlation of COMLEX-USA Level 2-PE with COMLEX-USA Level 1-CE (0.08) and the modest correlation with COMLEX-USA Level 2-CE (0.29) seem appropriate. The scores on the COMLEX-USA Level 1 predominately reflect understanding and knowledge of basic science concepts emphasized in the first 2 years of the curriculum, and therefore were not expected to correlate with clinical skills proficiency. For the clinical years, the COMLEX-USA Level 2-CE is designed to allow students to demonstrate knowledge of clinical concepts and principles involved in problem solving, but does not evaluate actual performance in a clinical setting. The modest associations between multiple-choice cognitive examinations and the clinical performance scores assessment provide evidence for the discriminant validity of COMLEX-USA Level 2-PE.

Components of WVSOM’s third-year OSCE had reliability coefficients ranging from 0.36 to 0.71, below the level of 0.80 that is generally acceptable for high-stakes examinations. However, a moderate correlation of 0.40 was found between the third-year OSCE Total Station score and COMLEX-USA Level 2-PE. Low performers on the third-year OSCE received detailed feedback and some additional instruction regarding their deficient areas, potentially homogenizing the scores and possibly attenuating the strength of this relationship. The school has since refined OSCE administration, achieving a Total Station score reliability of 0.85 in its April 2005 administration, so further research may result in a stronger association. Furthermore, passing the COMLEX-USA Level 2-PE was not a graduation requirement for the class of 2005; consequently, some students may not have taken the examination seriously.

The generalizability of this study is limited because the sample included only students from a single college of osteopathic medicine. Additional investigation of other variables is also warranted, including an analysis of the extent to which information available at the time of admission to medical school relates to COMLEX-USA Level 2-PE performance; the association of COMLEX-USA Level 2-PE scores and performance in osteopathic graduate medical education programs; and, to the extent that appropriate outcome measures can be identified, actual performance as a practicing osteopathic physician.

The NBOME has advised colleges of osteopathic medicine that in the future, when a student fails the COMLEX-USA Level 2-PE, the colleges will be provided with diagnostic information about a student’s performance (which students themselves received in 2004–2005), including which dimension was failed (J.R. Gimpel, DO, MEd, oral communication; August, 2005). Thus, colleges of osteopathic medicine will be able to provide more effective remediation for students and will conduct better statistical analysis regarding examination performance.

With the current national interest in ensuring the quality of practicing physicians and in reducing errors in the provision of medical care, the COMLEX-USA Level 2-PE has become an important measure in the licensing of osteopathic physicians. This external check on skills and abilities provides additional reassurance to the public that new osteopathic physicians are qualified not only in knowledge and cognitive skills, but also in interviewing, hands-on assessment, osteopathic manipulative medicine, and physician-patient communication and professionalism. While further research is certainly needed, the current study found appropriate relationships between school-generated academic variables and performance on the COMLEX-USA Level 2-PE, and therefore supports the validity of the COMLEX-USA Level 2-PE for assessing the clinical skills of future osteopathic physicians.

References


