COMLEX-USA and In-service Examination Scores: 
Tools for Evaluating Medical Knowledge Among Residents

Susan C. Sevensma, DO
Gerri Navarre, MSW
Robert K. Richards, PhD

From the Medical Education Department at Metro Health Hospital in Grand Rapids, Mich.

Address correspondence to Susan C. Sevensma, DO, Metro Health Hospital, Medical Education Department, 1919 Boston St SE, Grand Rapids, MI 49506-4160.

E-mail: susan.sevensma@metrogr.org.

Submitted August 17, 2007; revision received November 26, 2007; accepted December 4, 2007.

In 2003, the American Osteopathic Association (AOA) approved the adoption of seven core competency standards for internship and residency accreditation in all specialties.1 Although five of the competencies are specialty-wide and outlined by the AOA, the first two competencies—“Medical Knowledge” and “Osteopathic Philosophy and Osteopathic Manipulative Medicine”—are specific to residents’ specialties. As outlined by the AOA, specialty colleges were required to define and integrate the first two core competencies into their basic training standards by July 2004.1 Furthermore, AOA-accredited postdoctoral programs were expected to “fulfill the new specialty college requirements for these two core competencies.”

In an effort to meet these requirements, the medical education staff at Metro Health Hospital in Grand Rapids, Mich, considered multiple methods for the assessment of the core competency “Medical Knowledge.” Based on the universal availability of results from the Comprehensive Osteopathic Medical Licensing Examination-USA (COMLEX-USA) and specialty college in-service examinations, these tests were selected as measurement tools. However, rather than accept their apparent validity, we questioned whether these standardized examination scores measured medical knowledge—or merely test-taking skill.

In allopathic medical education, Ambroz and Chan2 reported on a 5-year retrospective study of 56 emergency medicine residents on the United States Medical Licensing Examination (USMLE) Step 1 and Step 2 and subsequent performance on their initial in-service examinations. According to the study,2 the scores had a linear correlation.

Another study3 demonstrated that scores on the Council of Resident Education in Obstetrics and Gynecology In-training Examination increased in each subsequent postgraduate year for all objectives and all years. The authors3 concluded that the examination was a dependable measure of the increase in residents’ cognitive knowledge.

Although studies on in-service examinations taken by osteopathic residents in emergency medicine, family medicine, general surgery, internal medicine, or obstetrics and gynecology are few and far between, several studies examined the relationship between COMLEX-USA scores and additional measures of student achievement.4-7 One study4 correlated scores on all three levels of COMLEX-USA with subsequent resi-
dent performance on in-service and board certification examinations. In that study, COMLEX-USA had high predictive validity for the American Osteopathic Board of Internal Medicine certifying examination results and internal medicine in-service examination scores.

As a result, we sought to further investigate the relationship between COMLEX-USA and in-service examination scores. More importantly, we sought to evaluate these tests’ usefulness as assessment tools for measuring residents’ medical knowledge.

**Methods**

From 2002 through 2006, COMLEX-USA and in-service examination scores for all interns and residents at Metro Health Hospital were gathered from medical education records and analyzed. The institutional review board at Metro Health Hospital reviewed our study protocol and approved the use of aggregate score data without obtaining individual consent.

All osteopathic medical graduates in their first year of postdoctoral training at Metro Health Hospital are referred to as residents. Therefore, for the purposes of the present study, the term *residents* encompasses interns and residents.

For our study population, we used residents’ final passing scores on all three levels of COMLEX-USA: Level 1, Level 2-Clinical Evaluation, and Level 3. Level 2-Performance Evaluation scores were not included.

To analyze the data, we divided residents into four score pattern categories:

- **A**—scores on all three levels of COMLEX-USA were at or above the national mean
- **B**—scores on two of the three levels of COMLEX-USA were at or above the national mean
- **C**—scores on one of the three levels of COMLEX-USA was at or above the national mean
- **D**—scores on all three COMLEX-USA levels were below the national mean

Using these four COMLEX-USA score pattern categories, we evaluated the in-service examination results of residents in the five specialties.

Osteopathic and allopathic residents in obstetrics and gynecology residency programs take the Council of Resident Education in Obstetrics and Gynecology In-training Examination. Osteopathic residents in emergency medicine, family medicine, general surgery, and internal medicine take annual in-service examinations created by their respective osteopathic specialty colleges, as follows:

- **Emergency medicine**—American College of Osteopathic Emergency Physicians
- **Family medicine**—American College of Osteopathic Family Physicians
- **General surgery**—American College of Osteopathic Surgeons
- **Internal medicine**—American College of Osteopathic Internists

Residents’ scores are then evaluated against the typical scores for the respective college’s in-service examination.

Individual residents took in-service examinations between one and four times, depending on the length of their training program and their postgraduate year at the start of the study. To test our hypothesis that valid measures of medical knowledge would show an upward progression in scores as residents advanced, we also studied trends in resident in-service scores for those who took the examinations at least twice during the study period.

An initial chi-square ($\chi^2$) analysis was used on all four score pattern categories to compare residents’ in-service examination scores above the mean versus those below the mean. Statistical significance for this test was considered a *P* value less than .05. Six additional $\chi^2$ analyses were run comparing all of the individual subgroups with one another to identify which differences were statistically significant. For these six subgroup analyses, statistical significance was defined as *P*<.008, demonstrating correlation between scores on two sets of examinations.

In 2005, one investigator (R.K.R.) met with each of the five program directors and asked them to name residents who had demonstrated outstanding medical knowledge since 2002. These names were obtained before the program directors were provided with our examination score analysis. Program directors were not asked to consider specific criteria in naming residents who excelled.

Also, to determine whether our study population was representative of all residents taking COMLEX-USA and in-service examinations, we compared residents’ scores with national COMLEX-USA and in-service examination scores.

**Results**

A total of 74 residents in the five specialties took 170 in-service examinations within the study period (Table 1). A total of 50 residents took between two and four in-service examinations. All of these residents showed substantial score increases, with 100% of internal medicine residents showing annual progress (Table 2).

Residents who consistently scored at or above the national mean on their COMLEX-USA examinations (category A) scored at or above the national mean on their in-service examination 70.7% of the time. Residents who consistently scored below the national mean on COMLEX-USA (category D) scored at or above the national mean on their in-service examinations only 20% of the time (Table 3).

Subgroup analyses revealed some statistically significant relationships. The greater number of test scores above the
national mean among residents who were in category A or B was statistically significant (P<.008) compared with the number of scores above the national mean among residents in category D. For residents in category A compared with those in category D, the statistical significance was even greater (P<.001). In addition, the number of test scores above the national mean for residents in category A were statistically significant compared with those in category C (P=.005).

Program directors in the five specialties named 24 residents who demonstrated outstanding medical knowledge. Of this cohort, 18 (75%) had COMLEX-USA category A or B score patterns and only 1 (4%) had a category D pattern.

In addition, residents in the present study had an average score of 524 for COMLEX-USA Level 1, 535 for Level 2, and 538 for Level 3. All scores were no more than 8% above the national COMLEX-USA mean of 500. The same residents’ scores on specialty in-service examinations met the national mean in the specialty by a range of 1.1% to 7%.

Comment
The findings in the present study that in-service examination scores generally correlated with COMLEX-USA score patterns are consistent with other studies. For example, Cavalieri found that “good performers” on written examinations. Correlation coefficients between COMLEX-USA scores and subsequent internal medicine examinations were significant and higher than .70. Similarly, Ambroz and Chan found that both USMLE Step 1 and Step 2 scores had a linear correlation with the emergency medicine in-service examination scores (P<.001).

However, such research is not without limitations. Cavalieri cautioned that specialty bias as a result of limiting research to a cohort of internal medicine residents could limit generalization of study findings to other disciplines. The present study differed in purpose, the number of specialties involved, grouping of residents, and focus on multi-year examination performance by residents in a single AOA-accredited institution. Nonetheless, even inclusion of five specialties at one institution does not permit generalization of our findings to all osteopathic specialties or other institutions. We continue to gather data on residents’ scores and explore collaborative studies to compare or contrast our initial results.

Another limitation of our study concerns how Metro Health Hospital program directors named residents who displayed outstanding knowledge. Although directors named individuals before they had access to our data, program directors may have been biased by their recollection of residents’ COMLEX-USA or in-service examination scores. In addition, we did not investigate the extent to which program directors
used COMLEX-USA Level 1 and Level 2 scores as a factor in selecting residents.

To minimize this limitation, we considered gathering data from Metro Health Hospital’s monthly resident performance evaluation forms to compare the naming of residents, but the form was revised in 2004 to better link faculty judgments with the core competencies. Therefore, we lacked a consistent format of faculty ratings of residents’ knowledge to compare with scores. In future studies, a consistent format should be used.

It is important to note that Metro Health Hospital does not exclusively use or advocate using COMLEX-USA scores as a single entity in the selection process for new residents. The examination’s three levels are intended to measure different aspects of medical knowledge and use different approaches to testing. Applicants who score below the national mean on Levels 1 and 2 may perform above the mean on Level 3. Even if they do not, some residents at Metro Health Hospital have earned above average in-service scores despite being assigned a category D COMLEX-USA score pattern.

We are exploring whether scoring below the national mean on COMLEX-USA Levels 1 and 2 suggest that special consideration be given to assess the medical knowledge of some residents during their first year at Metro Health Hospital. If residents’ scores on Level 3 and their first in-service examinations are below the national mean, additional educational opportunities to assess and improve residents’ medical knowledge may be needed.

In addition, the present study takes only one of the seven core competencies into consideration. Although “Medical Knowledge” is very important, all seven competencies must be taught and evaluated to make sure that a well-trained, competent, compassionate physician is the outcome of any training program.

Conclusion
In-service examination scores among Metro Health Hospital residents were generally correlated with COMLEX-USA score patterns and representative of all residents who took the examinations. As such, COMLEX-USA and specialty in-service examination scores are valid and useful tools to evaluate residents’ medical knowledge.

Acknowledgments
We thank Alan T. Davis, PhD, at the Grand Rapids Medical Education and Research Center for Health Professions for the statistical analyses used in this article, and William Cunningham, DO, Executive Vice President and Chief Medical Officer at Metro Health Hospital, whose consistent support for medical education and research led to this research project.

References