The purpose of this study was to evaluate the predictive validity of osteopathic medical licensing examinations for osteopathic medical knowledge measured by graduate written medical examinations. Performances on the three osteopathic initial licensing examinations, the three osteopathic internal medicine in-service examinations, and the osteopathic internal medicine board certification examinations were analyzed for a cohort of the most recent osteopathic internal medicine board certification examination candidates (N = 82). Multiple regressions were performed for the predictive value of licensing examination scores for the late examination scores. Logistic regressions were used for the prediction of pass/fail status on the licensing examinations for that on the board certification. A longitudinal performance profile was constructed to assess the rank changes in decile on the examinations at different times.

All correlation coefficients between the licensing examinations and other examinations were significant and higher than .70. The licensing examinations together predicted at least 60% of the variance of any of the other examination scores. The pass/fail status on the licensing examinations predicted 89% of the pass/fail status on the certification examination. Decile ranks since the first licensing examination were consistent and stable over at least 5 years.

The osteopathic licensing examinations had high predictive validity for the late written osteopathic internal medicine examinations. Generalization of the findings to other disciplines needs a caution, as a specialty bias may exist for these types of studies.

The responsibility of medical licensing examinations is to assure the public that candidates who pass the examinations have the minimal competence to practice medicine. This licensure concept implies a prediction of candidates’ status at the time they take the examinations for their status when they practice medicine. Accordingly, validation of the medical licensing examinations needs to determine how well the scores on the licensing examinations predict future performance. The written medical licensing examinations measure medical knowledge, cognitive skills, and judgment. Passing scores on written medical licensing examinations confirm candidates’ possession of medical knowledge, skill, and judgment necessary for entry-level practice. Although the ultimate test of validity is the consistency between performance on the licensing examinations and the outcomes of clinical practice, the predictive value of the scores on the written licensing examinations for the scores on future written examinations that assess a similar domain of medical knowledge and skills constitutes an important piece of validity evidence for the medical licensing examinations.

The purpose of this study was to evaluate how well performance on the osteopathic medical licensing written examinations predicted later performance on the major graduate written osteopathic medical examinations, ie, in-service and board certification examinations. The relationship between licensing examination results and clinical performance rating during residency and in actual practice involves complicated methodology issues and, therefore, is not addressed in this study.

All existing studies in this area are related to allopathic medical licensing examinations, as no research on the predictive validity of osteopathic medical licensing examinations exists. While a few authors have reported negative findings,1,2 most existing research has found a positive relationship between allopathic medical licensing examination scores and board certification examination results. Some of
the studies involved all three licensing examinations, while others involved two licensing examinations or one licensing examination. Some also investigated the relationship between the licensing examination scores and pass/fail status on board certification examinations. Those studies found that low licensing examination scores were associated with greater likelihood of failing board examinations. The relationship between licensing examination scores and in-service examination scores were much less studied, and findings on this topic were mixed. Analytical approaches for the studies in this area were mainly Pearson correlations and linear regressions.

Our study explored the relationship between the National Board of Osteopathic Medical Examiners’ (NBOME) licensing examinations and the subsequent graduate osteopathic medical written examinations. Two methodologic features separate this study from the previous studies. First, instead of selecting a single group of examinees, this study tracked a cohort’s performance from the first licensing examination (at the end of the second year of medical school) to the board certification examination (at the end of their residency training). Second, this study involved all three examinations in the licensing examination series. The inclusion of all three licensing examinations and all major graduate written examinations allowed for a more comprehensive investigation of predictive validity.

Methods Participants
The participants in this study were 82 of 132 first-time takers of the American Osteopathic Board of Internal Medicine (AOBIM) certifying examination in 2000. The 82 participants were chosen because they had taken both the Comprehensive Osteopathic Medical Licensing Examination–USA (COMLEX–USA) level 2 and level 3 before taking the AOBIM examinations. Examinees who took the Part II or Part III, the predecessors of the COMLEX–USA examinations, were not included in this study.

Instruments
The COMLEX–USA is a three-level licensing examination for osteopathic physicians designed to assess the knowledge component of the minimum competence of osteopathic medicine. All levels are constructed according to two dimensions: clinical presentations and physician tasks. While the overall complexity progresses from level to level, level 1 emphasizes scientific understanding of health and disease mechanisms, level 2 emphasizes diagnosis, and level 3 emphasizes management. Each level had 800 multiple-choice items. The reliability of COMLEX–USA examinations was in the range of .91 to .97. The COMLEX–USA standard scores were scaled to have a mean of 500 for all three levels and a standard deviation of 72 for level 1, 85 for level 2, and 111 for level 3. To pass level 1 or level 2, a standard score of 400 or more was required, while a score of 350 or more was required to pass level 3. All examinations were equated within levels so that scores from different administrations were statistically on the same scale and, therefore, comparable. The failing rates for levels 2 and 3 were about 12% and 9%, respectively.

The level 3 and level 2 examinations were implemented in 1995 and 1997, respectively. Level 1 was introduced in 1998 when participants in this study were already sitting for the level 2. Consequently, all participants completed the part I examination—the predecessor of the level 1—rather than the level 1 examination. The part 1 examination was a 900-item multiple-choice examination equally represented by seven basic science disciplines. The reliability of the part I examination was approximately .96, and an equated standard score of 400 was the minimum part I passing score. The failing rates of part I in its final 3 years were approximately 5%. Although both part I and level 1 test knowledge of basic sciences, the level 1 examination is clinical problem–oriented and, therefore, more clinical than the part I examination, which was discipline-oriented and more academic.

The American College of Osteopathic Internists (ACOI) offers a series of three annual in-service examinations for first-, second-, and third-year internal medicine residents to assess their progress during residency training. All examinations consisted of 200 multiple-choice items. The examinations did not designate pass or fail status. The examinees receive score reports in both raw score and Rasch-model based logit scores.

The AOBIM certifying examination was administered after 3 years of internal medicine residency training. It had 600 multiple-choice questions and was administered over 2 days. The first day consisted of two sessions, each with booklets of 165 items, while the second day had one session comprising 270 items. Candidates received the pass/fail outcome along with two-digit scaled scores converted from Rasch model–based logit scores. The Rasch model is an item-response theory model. The scores generated from this model are a logarithm of probability scores.

Procedures
The participants in this study took all seven written examinations (from part I to the board certification examination) in no less than 5 years. Participants’ COMLEX–USA scores were taken from the NBOME database and paired with the in-service examination scores and the board examination scores. If a participant had more than one score on any of the examinations, only the first-time score was used.

The metrics of part I, level 2, and level 3 used by this analysis were the three-digit standard scores as reported to candidates. For the in-service examinations, logit scores were chosen for analysis, as they had better statistical properties than raw scores. The metric of the AOBIM certifying examination used in this study was the logit score, as the scaled scores were initially converted from the logit scores.
The predictive value of COMLEX–USA for the prospective written examination performances was investigated from two aspects: statistical predictive value of COMLEX–USA scores for graduate examination scores in general and statistical predictive power of COMLEX–USA pass/fail status for the pass/fail status on the board certification examinations in specific. For the first topic, this study first examined pairwise Pearson correlation coefficients among part I, levels 2 and 3, and graduate written examinations as a descriptive analysis. Further, multiple linear regressions were performed separately for each in-service examination and the board examinations, with part I, level 2, and level 3 scores as predictors. Two modeling approaches were used: a model including all three licensing examination scores as predictors and a model selected by stepwise approach.

For the second topic, this study first used the Mantel-Haenszel test to examine the associations between pass/fail status on the certification examination and those of each of the licensing examinations. The Mantel-Haenszel method tests the significance among variables with categorical data. Further, logistic regression analyses were used with the pass/fail status of part I, level 2, and level 3 as predictors to predict the pass/fail status on the AOBIM certifying examination. Eight logistic regression models were fit. The first model served as a baseline, which did not use any of the licensing examinations. The next three models used only one of the three licensing examinations as the predictor. Models 5, 6, and 7 used only two of the three licensing examinations as the predictors. Model 8 used all three licensing examinations. The best predictive model was selected by a series of goodness-of-fit tests.

The participants in this study took all seven written examinations (from part I to the board certification examination) in no less than 5 years. Although comparing seven scores on those examinations was not appropriate because those examinations were not in the same measurement scale, comparing each participant’s ranks over time can provide some insight of the predictive value of licensing examinations. The rank comparison of this study was represented by a plot of participants’ decile ranks on the seven examinations.

Results
Initial diagnostic analyses found that one participant had high scores on all three licensing examinations but a low score on the board certification examination. Statistically, that participant was an outlier in the relationship between each of the licensing examinations and the certification examination. Although inclusion of these data had no adverse affect on the outcome of analyses, this participant was excluded from the reported analyses because the scores were nonrepresentative of the data and did not have an appreciable effect on the results.

Descriptive statistics (Table) suggested that this group of participants in general was not different from the COMLEX–USA candidate population. The mean score of the part I candidate population for the past 3 years that examination was used was approximately 540, with a failing rate of 5%. The mean score on part I for this group of participants was better, but the failure rate of this group (8.6%) was higher. The par-

<table>
<thead>
<tr>
<th>Examination</th>
<th>N</th>
<th>Mean (SD)</th>
<th>% Fail</th>
<th>PI</th>
<th>L2</th>
<th>L3</th>
<th>IS1</th>
<th>IS2</th>
<th>IS3</th>
<th>AOBIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI</td>
<td>81</td>
<td>562 (105)</td>
<td>8.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2</td>
<td>81</td>
<td>507 (86)</td>
<td>11</td>
<td>.783</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L3</td>
<td>81</td>
<td>544 (139)</td>
<td>6.2</td>
<td>.816</td>
<td>.871</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS1</td>
<td>76†</td>
<td>.083 (.624)</td>
<td>NA</td>
<td>.770</td>
<td>.805</td>
<td>.799</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS2</td>
<td>81</td>
<td>.302 (.658)</td>
<td>NA</td>
<td>.750</td>
<td>.717</td>
<td>.757</td>
<td>.921</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS3</td>
<td>81</td>
<td>.623 (.679)</td>
<td>NA</td>
<td>.706</td>
<td>.753</td>
<td>.774</td>
<td>.895</td>
<td>.917</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AOBIM</td>
<td>81</td>
<td>.677 (.607)</td>
<td>8.6</td>
<td>.716</td>
<td>.749</td>
<td>.759</td>
<td>.800</td>
<td>.809</td>
<td>.850</td>
<td></td>
</tr>
</tbody>
</table>

*All correlation coefficients were significant at the .05 level.
†Five participants had not completed this in-service examination during their first year for unknown reasons but were included in all other analyses.
PI indicates part I, the predecessor of the Comprehensive Osteopathic Medical Licensing Examination–USA level 1; L2, Comprehensive Osteopathic Medical Licensing Examination–USA level 2; L3, Comprehensive Osteopathic Medical Licensing Examination–USA level 3; IS1, American College of Osteopathic Internists’ in-service examination for first-year internal medicine residents; IS2, American College of Osteopathic Internists’ in-service examination for second-year internal medicine residents; IS3, American College of Osteopathic Internists’ in-service examination for third-year internal medicine residents; AOBIM, American Osteopathic Board of Internal Medicine certification examination.
participants’ performance on level 2 was almost identical to the level 2 national mean of 500. The level 3 mean performance was similar to the level 3 national mean of 520, but the failing rate was lower than that of the level 3 population’s 9%. For part 1 and level 2, the variation of scores in the samples was similar to the candidate population of each examination. The variation of this group’s level 3 performance was larger than the population variation of this examination.

The Table also shows the intercorrelations among all the examinations. Not surprisingly, the three licensing examinations correlated highly with each other. The correlations of the certification examination with levels 2 and 3 were .75 and .76, respectively, and the correlation with the part I was .71. Most of the correlations of licensing examinations and in-service examinations were greater than .75, though correlations with the part I were lower.

The adjusted R²s of the multiple linear regressions with all licensing examination scores as predictors indicated that the three licensing examinations could explain 61% of the variation of certification examination results. They also explained 70%, 62%, and 62% of the three in-service examination performances. Stepwise regression results suggested that levels 2 and 3 could sufficiently predict 60% of the variation of certification examination performance. Adding part I into the equation did not statistically improve the prediction. Similarly, part I and level 2 predicted 69% of the in-service 1 examination outcome variation, with the level 3 excluded as a non-significant predictor. For the in-service 2 examination, part I and level 3 explained 62% of the outcome variation. For the in-service 3 examination, levels 2 and 3 were the significant predictors, explaining 62% of the variation.

Mantel-Haenszel chi-square tests of the associations between the pass/fail status of licensing examinations and that on the certification examination found that the pass/fail status on all licensing examinations was significantly associated with the status on certification examination.

In predicting pass/fail on the certification examination by logistic regressions, the full model, which used all three licensing examinations as predictors, and the model 7, which used levels 2 and 3 as predictors, were the best. Nevertheless, the improvement of the full model to the model 7 was not significant. The model 7’s overall correct prediction rate was
89%. For 72 of the 81 participants, pass/fail status on the certification examination was correctly predicted by the model. However, the model tended to predict passing status more accurately than failing status. The correct prediction for passing was 98.6%, and the correct prediction for failing was 11.1%. Obviously, this was due to thin data of failing status on both licensing examinations and the certification examination.

The Figure depicts participants' decile rank changes on the seven examinations over a minimum 5-year period using the rank on part I as the baseline. Clearly, mean decile ranks were consistent in general over time. As the Figure demonstrates, participants in five decile groups did not change ranks by more than .5 decile. Participants in two decile groups changed ranks by 1 to 1.5 deciles. Three groups changed ranks by approximately 2 deciles. Those were the groups of the lowest, the middle, and the highest deciles at the beginning of the profile.

Discussion
The outcome of all analyses suggested that the osteopathic medical licensing examinations consistently and strongly predict consequent osteopathic medical written examination performances over time. The correlations of all licensing examinations with other written examinations were high, and the predictions were strong. Although technically it was not appropriate to compare the predictive power for different examinations, it was remarkable that the licensing examinations held substantial and consistent predictive power for performance on similar examinations over time. Another noticeable aspect was that the part I examination was slightly less associated with other examinations and contributed little to the prediction in the late stage of residency. To some extent, this was expected because the part I examination tested basic sciences in a relatively less clinical fashion, and other examinations (levels 2 and 3 and the graduate written examinations) were clinically oriented. One feature of this study is that its sample size was small. Usually, when the sample size is larger, the score range is wider and the correlation and regression results are better. When sample size is small, the opposite occurs. Considering the small sample size of this study, the outcomes appeared convincing.

Results of the pass/fail status analyses were consistent with the results from correlation and regression analyses. Overall, status on the licensing examinations was associated with and predicted the status of the board certification examination. This association or prediction for those who passed the certification examinations was strong. The prediction of failing the board examination was weak due to the absence of more participants who failed the examination. Given the robust overall relationship between the pass/fail status of the two examinations in this study, COMLEX-USA’s predictive value for the failure status on certification examinations can be better demonstrated if the data of failing status are sufficient.

The decile change profile graphically confirmed the findings of other analyses. The intention of this profile approach was not to mask the reality that individuals’ ranks could vary substantially. Rather, we used this approach to detect a possible general pattern. As shown, overall, good performers in the licensing examinations as grouped in decile ranks remained good performers throughout the postgraduate years, or vice versa, as far as written examinations were concerned.

Our study should be interpreted with caution, however. This study only dealt with the predictive validity of the osteopathic medical licensing examinations for the knowledge component of the medical competency as measured by written examinations. The question of how the licensing examinations predict future clinical performance—the more critical aspect of the validation of medical licensing examinations—was not answered and needs more devoted research efforts.

Although this study demonstrated that the osteopathic medical licensing examinations predicted future performance on written medical examinations well, it must be emphasized that the purpose of osteopathic medical licensing examinations was not to predict candidates’ performance in residency or success on certification examinations. Therefore, it was not this study’s intention to recommend residency programs to use licensing examination scores to select or evaluate residents, or use those scores to predict residents’ scores on board certification examinations.

The results of this study were based on participants in a single discipline. Caution is also needed when generalizing the results of this study to disciplines other than internal medicine. As Gonnella et al11 suggested, the association of licensing examination and board examination scores are high for students in disciplines such as internal medicine. Thus, a specialty bias may exist for these types of studies. As all performances in this study were written examination outcomes, there was a possibility that the reason for a good association of two written examinations was not that they measured the same construct, but that they needed the same test-taking skills. Because this study solely examined written examinations, it is possible (though difficult to prove) that some of the predictions were accurate because they predicted test-taking skills and not actual medical knowledge.

Overall, this study provides strong evidence for the predictive validity of osteopathic medical licensing examinations for written examinations. However, the most critical evidence of the validity of medical licensing examinations is the performance of the licensed physicians in clinical practice. More research efforts are needed in this direction.

Acknowledgment
The authors thank Corey Hammer, MA, for his assistance in preparing and analyzing data and for editing this article.
MEDICAL EDUCATION

References


