Special Committee on Licensing Examinations

Federation of State Medical Boards of the United States, Inc.

The recommendations contained herein were adopted as policy by the House of Delegates of the Federation of State Medical Boards of the United States, Inc., April 2001

Section I: Background

In May 1998, then President William H. Fleming, III, MD appointed the Special Committee to Evaluate Licensure Examinations to implement Resolution 98-1, adopted by the Federation House of Delegates at its 1998 Annual Business meeting. The resolution requested the Federation of State Medical Boards meet with the National Board of Medical Examiners (NBME), the National Board of Osteopathic Medical Examiners (NBOME), and any other interested parties to document the validity of the United States Medical Licensing Examination (USMLE) and the Comprehensive Osteopathic Medical Licensing Examination (COMLEX)-USA and to provide that information to the House of Delegates at its 1999 Annual Meeting. The resolution also requested that the House of Delegates consider its position on examinations for licensure to include recognition of COMLEX-USA.

The committee met three times between September 1998 and February 1999. In February 1999, the NBOME notified the committee that it was temporarily disengaging from the initiative in order to conduct its own studies documenting the validity of the COMLEX-USA. The committee subsequently issued a final report that was adopted by the House of Delegates at its 1999 Annual Meeting. The report contained the following recommendations:

- The FSMB House of Delegates reaffirm its policy that USMLE be the single pathway to licensure for all U.S. physicians.
- The FSMB reappoint a special committee to reevaluate the COMLEX-USA and the USMLE at such a time as NBOME is able to provide appropriate validity studies regarding the COMLEX-USA.

In April 1999, representatives from the NBOME notified the Federation of their intent to reengage with the committee. In keeping with the recommendation issued by the House of Delegates, Federation President Alan E. Shumacher, MD, appointed the Special Committee on Licensing Examinations, to complete the remainder of the charge and report back to the Federation's House of Delegates as called for in Resolution 98-1.

For the remainder of this report, the term "committee" refers collectively to both the Special Committee to Evaluate Licensure Examinations and the Special Committee on Licensing Examinations.

Section II: Objectives

The committee was charged with the following objectives:

1. Identify the desired scope of an examination used to determine qualification for initial physician licensure.
2. Identify the critical elements used to design, develop, administer and score a licensure examination.
3. Evaluate the validity and reliability of the critical elements used in the licensure examination.

4. Evaluate data to determine the extent these critical elements are employed in the construct of the USMLE.
5. Evaluate data to determine the extent these critical elements are employed in the construct of the COMLEX-USA.
6. Determine to what extent each examination meets its stated purpose.
7. Identify variances in the design, development, administration and scoring of the USMLE and the COMLEX-USA and determine to what extent these variances affect the comparability of the examinations.
8. Develop a recommendation to the Federation membership as to whether the COMLEX-USA should be endorsed as a separate but comparable pathway to the USMLE for the purpose of licensing osteopathic physicians.

In discussing its charge, the committee agreed to look at both the validity and the comparability of the two licensure examinations.

Section III: Elements Critical to the Development of an Examination Program
The committee recognizes that test development and validation practices are important components in documenting the validity of any medical licensure examination. Stephen G. Sireci, PhD, psychometric consultant, was hired by the committee to evaluate the test development and validation practices of the USMLE and the COMLEX-USA. Dr. Sireci is a nationally noted psychometrician who both the NBME and NBOME agreed was qualified and appropriate for the work of the committee. Dr. Sireci had no prior nexus with any of the organizations involved in this study.

The committee identified the following critical elements as being important in determining the validity and comparability of the COMLEX-USA and the USMLE:

Purpose:
Broad in scope.
Should not focus on any one discipline or specialty.

Design:
Incorporates all aspects of basic scientific principles necessary to practice medicine.

Adequately measures physician skills, tasks and clinical encounters in the appropriate setting (i.e. office practice, satellite health center, emergency services and in-patient facilities).

Development Process:
Steps and processes used to construct the licensure examination are appropriate.

Sufficient number of committees, subcommittees and task forces to adequately produce the number of test items necessary to maintain an adequate pool of questions and number of test forms for the examination population.

Physician licensure/practice community participation in the development and validation of test items, thereby ensuring adequate representation from the medical practice community across all disciplines.

Current and ongoing enhancements designed to further strengthen the development process of the examination.

Scoring:
Methods used in determining test score reliability and the reliability of pass/fail decisions.

Types of methods used to determine the standards for performance.
How the Pass/Fail standards are developed and executed over time.

Research and Development:
Surveys of candidates and licensing authorities to evaluate the strengths and weaknesses of the examination program.

Documentation of validity studies conducted to appraise the utility and efficiency of the examination program and that support the use of the exam scores for the purpose of licensing physicians.

Research and other activities focused on improving the examination program, such as computer-based testing and performance-based assessment.

Administration:
Parent organizations provide sufficient staff support, infrastructure, and resources required to maintain a quality examination that represents the high standards expected by the public and medical profession.

Examination is "open" to all examinees that meet the licensure requirements.

The examination administration provides for adequate physical security to prevent compromising the integrity of the examination and the licensure process (security would also extend to providing an adequate number of test items and forms to guard against memorization, etc.)

The examination and exam process is independent and free to operate without undue influence by organizations that may influence the content, standards and purpose of the examination.

Additionally, Dr. Sireci asked committee members to complete a task designed to obtain their opinions about (1) the differences and similarities among the stated purposes associated with each examination, (2) the importance of the content areas within the examinations, (3) the similarities and differences among the content specifications for the two examinations, and (4) the relative appropriateness of the COMLEX-USA and the USMLE for initial physician licensure. The results indicated committee members perceived that a medical licensure examination should test minimal competency to ensure protection of the public. Furthermore, there was consensus that the examinations' purpose statements are appropriate and that both exams contain important content areas. Finally, the results of the content comparison portion of the task showed a large degree of overlap across the two examination programs.

Section IV: Review of Technical Documentation
At the committee's request, the NBME and the NBOME provided the following technical documentation to Dr. Sireci for evaluation: (1) complete descriptive statistic and item analysis information, (2) documentation of the test development process, including how the passing scores are set and how items and test forms are reviewed (internally and externally) before becoming operational, (3) model fit or dimensionality analyses, (4) item bank information, (5) validity studies and (6) any other information describing the quality of their testing program. An electronic search was also conducted (using ERIC, MEDLINE, and PSYCHINFO) to discover other papers and publications that were pertinent to an evaluation of the validity argument for either examination.

The psychometric evaluation focused on evidence that would support the argument that each examination is valid for its stated purpose. A "validity argument" is a body of evidence that supports the use of a test for a particular purpose and provides evidence that decisions made on the basis of test scores are logical, defensible and fair. The criteria used by Dr. Sireci to evaluate these validity arguments were:

1. appropriateness of test development procedures (development of test specifications, item development, content domain representation, item quality, sensitivity review, and other content related evidence),
2. appropriateness of standard setting procedures (reasonableness of standard setting method, selection of participants, technical adequacy of the procedure, reliability of the standards, validation of the standards) and
3. evidence supporting the validity of score-based inferences (test score reliability analyses, passing score reliability analyses, dimensionality analyses, differential item functioning analyses, test bias analyses, analysis of group differences, other construct validity studies).

Other criteria for evaluating validity arguments that were not emphasized in the review included scaling, equating, test administration, test security, test accommodations and score reporting. **Attachment 2** contains a detailed description of the criteria used to evaluate the validity arguments for the COMLEX-USA and the USMLE.

After reviewing all materials, each examination program was given one of four possible ratings for each of the three categories listed above. The four rating categories were: insufficient, marginal, satisfactory, and exemplary. Finally, an overall conclusion was provided regarding whether there was sufficient evidence to support the validity argument for each examination. The examinations received the following ratings:

USMLE

Appropriateness of Test Development Procedures
Satisfactory

Appropriateness of Standard Setting Procedures
Exemplary

Evidence Supporting the Validity of Score-Based Inferences
Exemplary

Conclusion: There is a strong body of evidence to support the USMLE validity argument.

COMLEX-USA

Appropriateness of Test Development Procedures
Satisfactory

Appropriateness of Standard Setting Procedures
Marginal

Evidence Supporting the Validity of Score-Based Inferences
Exemplary

Conclusion: There is sufficient evidence to support the validity argument for the COMLEX-USA.

The committee noted that the process used to set the passing level for the COMLEX-USA Levels 1 and 2 was reasonable and defensible. However, due to the low number of panelists used and the lack of a follow-up study in the last five years for Level 3, the committee concurred with its consultant that a marginal rating for appropriateness of standard setting procedures was merited.

Following review of its psychometric consultant's evaluation of the technical documentation submitted by the NBME and NBOME and review of articles identified in the electronic database search, the committee

found that both the USMLE and the COMLEX-USA are valid for their stated purpose. **Attachment 3** contains Dr. Sireci's report on his evaluation of both examination programs.

As a final step in its review of the USMLE and COMLEX-USA, the committee viewed content from actual test forms for both examination programs.

Section V: Determining Comparability
In order to develop a recommendation regarding endorsement of the COMLEX-USA as a separate but comparable pathway to licensure for osteopathic physicians, the committee determined it needed to evaluate the comparability of the two examination programs in the areas of quality of content and rigor of standards. Because such studies do not currently exist, the committee directed its independent consultant to develop proposals outlining ways in which the necessary data could be obtained. In discussing the proposals developed by Dr. Sireci in cooperation with NBME and NBOME (**Attachment 4**), the committee voiced concerns as to whether a study comparing the examinations' degree of difficulty could reasonably be conducted in a way that would result in a valid outcome. Concerns were also voiced about the expense and time associated with the studies proposed to date. As a result of this discussion, the committee determined it was unable to document the comparability of the two examinations in the areas of rigor and quality.

The committee further agreed that it would discontinue pursuit of comparability studies in light of the commitment articulated by the Federation, the NBME and the NBOME to pursue discussions regarding the development of a common examination system.

Section VI: Committee Findings
The committee believes its most critical responsibility is to develop a policy recommendation that will assist medical boards in protecting the public by ensuring that all physicians applying for initial licensure, whether allopathic or osteopathic, have obtained the knowledge and skills that are appropriate for the unsupervised practice of medicine and that meet a single, acceptable standard for licensure. The committee acknowledges that one way to accomplish this may be to recognize the COMLEX-USA as a separate but comparable pathway to licensure for osteopathic physicians. However, there are no data demonstrating the USMLE and the COMLEX-USA are comparable in rigor or quality.

The committee believes strongly that the Federation should take a leadership role in working with NBME and NBOME to develop and offer a pathway for licensure that is truly uniform. Therefore, the committee recommends that the Federation:

1. Recognize that the USMLE and the COMLEX-USA are valid for their stated purposes;
2. Affirm the intent behind the Federation's policy regarding a single pathway to licensure; and finally,
3. Collaborate with the NBME and the NBOME to establish a concrete plan toward the development of a common examination system to assure the public that all physicians are meeting a uniform standard for purposes of medical licensure.

Respectfully submitted,

Special Committee on Licensing Examinations
Alan E. Shumacher, MD, Chair
George C. Barrett, MD
William H. Fleming, III, MD
Hartwell Z. Hildebrand, MD
N. Stacy Lankford, MD
D. Jayne McElfresh
Thomas R. Pickard, DO
R. Russell Thomas, Jr., DO, MPH

Special Consultant
Stephen G. Sireci, PhD

Special Committee to Evaluate Licensure Examinations
William H. Fleming, III, MD, Chair
Robert Foster, DO
N. Stacy Lankford, MD
William M. Lightfoot, MD
D. Jayne McElfresh
Thomas R. Pickard, DO
Alan E. Shumacher, MD
Stephen I. Schabel, MD
R. Russell Thomas, Jr., DO, MPH
Charles E. Trado, Jr., MD

Special Consultant
Stephen G. Sireci, PhD

The Federation would like to recognize the following individuals for their involvement in and commitment to the work of the committee:
Thomas Cavalieri, DO, NBOME
Donald Melnick, MD, NBME
Frederick Meoli, DO, NBOME
Ronald Nungester, PhD, NBME
Stephen Sireci, PhD
Joseph Smoley, PhD, NBOME

The Federation would like to extend special gratitude to Stephen G. Sireci, PhD, for his assistance and expertise in developing information regarding the technical evaluation of the USMLE and COMLEX-USA programs. The Federation is also indebted to the individual committee members for their invaluable contributions and service to this project.

Attachment 2

Validation Process
Documenting and evaluating the validity of a licensure examination is a difficult endeavor. The process is complex and cannot be reduced to a single statistical test or checklist of steps completed. The most widely cited guidelines for evaluating licensure examinations are the Standards for Educational and Psychological Testing, which were developed jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1985). Many psychometricians who were instrumental in developing validity theory and designing validation processes contributed to this document (e.g., A. Anastasi, L.J. Cronbach, R. Linn, S. Messick, M. Novick, L. Shepard, etc.), and it has been used by the courts in litigation surrounding employment and certification testing. As stated in its introduction, "the purpose of publishing the Standards is to provide criteria for the evaluation of tests, testing practices, and the evaluation of test use" (p. 2). This document was used to derive criteria for evaluating the COMLEX-USA and USMLE. Other seminal works in this area were also used to derive evaluation criteria.1[1] Several important facets of test validity are described in these documents. Some fundamental aspects of test validity and validation to be understood from the outset are:

1[1] Other seminal works describing test validity and test validation that were consulted were Messick (1989), Wainer and Braun (1988), Kane (1992), and Shepard (1993). Research specific to licensure and

- Tests cannot be valid or invalid. Tests must be evaluated with respect to a particular purpose. It is the decisions (inferences) made on the basis of test scores that are validated, not the test itself.
- Evaluating inferences made from test scores involves several different types of qualitative and quantitative evidence. The process of test validation is similar to providing evidence in support of an argument. Specifically, the validity argument is that test scores are valid for making specific inferences about test takers.
- Evaluating the validity of inferences derived from test scores is not a one-time event; it is a continuous process. Therefore, test developers and test users must continually seek to gather evidence in support of the validity argument.
- Test developers must provide evidence to support the validity of inferences derived from their test scores. Test users must evaluate this evidence to ensure it supports their purposes for using the test.

Although the Standards point out that "evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every primary standard in this document" (p. 2), there are several minimal requirements that all testing programs must satisfy. For example, "the Standards advocates that, within feasible limits, the necessary technical information be made available so that those involved in policy debate [regarding the test] be fully informed" (p. 1). The Standards also provide important advice for evaluating tests: "It would not be appropriate for test developers or test users to state that a test, manual, or procedure satisfies or follows these standards. That judgment is more appropriately made by others in the professional community" (p. 2).

In the spirit of such advice, several standards were taken as requisite for a quality professional licensure examination for the medical profession.2[2] For example,

Standard 3.1: Tests and testing programs should be developed on a sound scientific basis. Test developers should compile the evidence bearing on a test, decide which information is needed prior to test publication or distribution and which information can be provided later, and conduct any needed research. (p. 25).

In addition, it should be mentioned that there are several standards that stipulate how technical material, such as validity evidence, should be made available to test users and consumers of test data. For example,

Standard 5.3: Technical manuals should cite a balanced and representative set of studies regarding general and specific test uses. The cited studies, except published work, dissertations, and proprietary studies, should be made available on request to the prospective test user by the publisher. (p. 36)

Standard 5.5: Test manuals should be amended, supplemented, or revised as necessary to keep information for users up-to-date and to provide useful, additional information or cautions. (p. 36)

Standard 5.6: In addition to publishing test manuals with amendments, supplements, and revisions, test developers and publishers should respond to reasonable requests for additional information, which may be required in order to determine the appropriateness of intended test use. (p. 36)

---

certification exams were also considered (Downing & Haladyna, 1997; Kane, 1994; Kuehn, 1990; Smith & Hambleton, 1990).

2[2] The specific standards from the AERA, APA, and NCME *Standards* (1985) emphasized in the evaluation of the COMLEX-USA and USMLE were: 1.1, 1.5, 1.6, 2.1, 2.10, 2.12, 3.1-3.5, 3.8, 3.10, 3.12, 3.13, 3.16, 4.1, 4.2, 5.1-5.3, 5.5, 5.6, 6.9, 11.1-11.3. These guidelines were summarized using three evaluation categories (appropriateness of test development procedures, appropriateness of standard setting procedures, and evidence supporting the validity of score-based inferences), as described in the Committee's evaluation report.

It is important to note that these standards require validity studies to be conducted, to be made available to interested parties, and to be ongoing. As the evaluation report describes, the vast difference between the validity evidence for the COMLEX-USA and the USMLE was the large number of studies conducted by the USMLE that were readily available upon request in contrast to the lack of studies regarding the COMLEX-USA. In addition, it was notable that the NBME conducted numerous studies on the USMLE's predecessor. These studies formed the basis for the USMLE validity research agenda, which began well before the USMLE became operational (e.g., La Duca, et al.,1984; Nungester et al., 1991).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, 10, 61-82.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112,527-535.

Kane, M. (1994). Validating interpretive arguments for licensure and certification examinations. *Evaluation & The Health Professions*, 17, 133-159.

Kuehn, P. A., Stallings, W. M., 7 Holland, C. L. (1990). Court-defined job analysis requirements for validation of teacher certification tests. *Educational Measurement: Issues and Practice*, 21-24.

LaDuca, A., Taylor, D. D., & Hill, I. K. (1984). The design of a new physician licensure examination. *Evaluation & The Health Professions*, 7, 115-140.

Nungester, R. J., Dillon, G., F., Swanson, D. B., Orr, N. A., & Powell, R. D. (1991). Standard setting plans for the NBME Comprehensive Part I and Part II Examinations. *Academic Medicine*, 66, 429-433.

Messick, S. (1989). Validity. In R. Linn (Ed.), Educational measurement, (3rd ed.). Washington, D.C.: American Council on Education.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405-450.

Smith, I. L., & Hambleton, R. K. (1990, winter). Content validity studies of licensure examinations. *Educational Measurement: Issues and Practice*, 7-10.

Wainer, H., & Braun. H. I. (Eds.), *Test validity*. Hillsdale, New Jersey: Lawrence Erlbaum.

Attachment 3

Psychometric Evaluation of the COMLEX-USA and USMLE
Report Submitted to the Federation of State Medical Boards
Special Committee to Evaluate Licensure Examinations
Stephen G. Sireci, Ph.D.
Independent Psychometric Consultant

This report summarizes the evaluation of the technical material submitted by the National Board of Medical Examiners (NBME) in support of the United States Medical Licensing Examination (USMLE), and the technical material submitted by the National Board of Osteopathic Medical Examiners (NBOME) in support of the Comprehensive Osteopathic Medical Licensure Examination (COMLEX-USA).

Definitions and Underlying Assumptions
Several psychometric publications related to the validity of licensure examinations were included with the Committee mailing for the December 2, 1998 meeting. These publications were used to help describe the theoretical foundation underlying the evaluation process and the criteria to be used in the evaluation. The most widely used document for determining whether test developers adhere to accepted standards of test quality are the Standards for Educational and Psychological Testing, produced by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1985). Although this entire document is important for considering issues of test validity, only excerpts from the validity chapter, and the chapter on licensure and certification testing, were reproduced for the Committee. Other publications reproduced were Smith and Hambleton (1990), Kuehn, Stallings, and Holland (1990), Downing and Haladyna (1997), and Kane (1994).

Six assumptions regarding the concept of "validity" were central to the evaluation of the technical information sent by the NBME and NBOME:

- Validity refers to the soundness, meaningfulness, and appropriateness of inferences derived from test scores.
- Inferences must be validated with respect to the testing purpose.
- What we seek to validate are decisions made on the basis of test scores.
- Validity is not "all or none." Tests are not "valid" or "invalid."
- Test validation is an ongoing process.

The evaluation was conducted from the perspective that to claim a licensure exam is valid for its stated purpose requires a "validity argument," which is a body of evidence supporting the use of a test for a particular purpose. A validity argument provides evidence that decisions made on the basis of test scores are logical, defensible, and fair.

Method and Criteria
Letters were sent to the NBME and NBOME requesting all materials related to their "test development and validation practices" (see Tab D of December 2, 1998 meeting materials). These materials were taken to represent the validity argument for each examination program. Three general criteria were used to evaluate the validity arguments:

Appropriateness of test development procedures.
This category includes data related to: the development of test specifications, item development, content validity studies, adherence to item writing guidelines and other indicators of item quality, sensitivity review, item selection criteria, score reliability analyses, and scaling procedures.

Appropriateness of standard setting procedures.
This category includes data related to the reasonableness of the standard setting method, selection of standard setting panelists, technical adequacy of the procedure, reliability of the standards, and validation of the standards.

Evidence supporting the validity of score-based inferences.
This category includes dimensionality analyses, differential item functioning analyses, test bias analyses, analysis of group differences, and any other construct validity studies.

In addition to the materials sent by each examination Board, electronic searches were conducted to discover other papers and publications that were pertinent to an evaluation of the validity argument for either exam. A detailed summary of the electronic database report was sent to the FSMB on December 11, 1998.

After reviewing all materials, each examination program was given one of four possible ratings for each of the three categories. The four rating categories were: insufficient, marginal, satisfactory, and exemplary. Finally, an overall conclusion was provided regarding whether there was sufficient evidence to support the validity argument for each exam.

Results
The results for the USMLE are presented first, followed by the results for the COMLEX-USA.

USMLE Results

USMLE: Appropriateness of Test Development Procedures

Development of test specifications
Three NBME studies were provided that described how the USMLE test specifications were developed (LaDuca et al., 1984; NBME, 1998; Willetts-Bloom et al., 1993). Four data sources were used to identify diseases and problems encountered by physicians: National Ambulatory Medical Care Survey (NAMCS), National Hospital Discharge Survey, Market Scan (from Medstat systems), and the National Disease and Therapeutic Index (from IMS-America). These data sources, and other survey data gathered by NBME, were used to develop a preliminary "practice model" for an undifferentiated (generalist) physician. Physician panels reviewed and revised the practice model and worked with the NBME staff to derive preliminary test specifications. Subsequent physician panels reviewed and approved the test specifications. The current test specifications reflect the consensus of these external review panels. These specifications undergo periodic external review to reconfirm their viability.

Development of test items
Test items are developed by committees of experts from a variety of medical specialties. A comprehensive item writers' manual was developed by NBME staff and appears to be an excellent training manual. External physician panels review test items and preliminary test forms. Across the three Steps, almost 500 external physicians have been involved in reviewing or writing tests, items, or test specifications. At least one D.O. is involved in the review of each Step.

our studies were provided that evaluated the utility of different item formats used or proposed for the USMLE. These studies guided selection of the best item formats. Sample test items are provided to candidates and these items adhere to accepted item writing guidelines.

Development of examinations
Appropriate item analysis procedures are used to select items and validate the item keys. Currently, items are pilot-tested and selected based on content characteristics and statistical characteristics (item difficulty and discrimination). Both classical and item response theory (IRT) item statistics are used to select items. Item statistics are also used to validate the keys and screen out problematic items from operational scoring. Classical item analyses include computation of p-values, conditional p-values, point biserial and biserial correlations, "high-low" statistics, and complete distractor analyses. IRT item analyses include item infit and outfit statistics. The average item biserials are adequate.

Other content-related evidence
Two examinee surveys were conducted to appraise test quality, and the examinees generally supported the quality of the exams. Also, there is extensive documentation for candidates and lay audiences describing

the test content, administration, and scoring. On the negative side, no content representation studies, nor sensitivity review studies, are conducted.

Scaling and scoring of examinations
All USMLE items are scored correct or incorrect. Each Step of the USMLE is scaled (calibrated) using the one-parameter (Rasch) IRT model. The fit of this model to the data is evaluated for each exam administration and conditional standard errors of measurement are reported along the entire score scale. Classical internal consistency reliability estimates (KR-20) are computed and are consistently high (above .90). The pass/fail decision (standard setting) is described below.

Rating: Based on evaluation of the above, the appropriateness of the USMLE test development procedures was rated "satisfactory."

USMLE: Appropriateness of Standard Setting Procedures
The documentation supporting the USMLE standard setting procedures included six research papers, one summary report, and four newsletter articles. The standard setting panelists are selected from a variety of disciplines and at least one D.O. sat on each panel. Efforts are made to achieve sex and racial/ethnic equity within the standard setting panel. For example, about 20-25% of the panelists for the Level 2 and 3 studies were non-Caucasian. The percentage of female standard setting panelists ranged from 17% to 45% across the most recent studies for each Step. Medical specialty area and geographic region are also used as criteria for selecting standard setting panelists.

A comprehensive multi-step process is used to set the standards. Two independent standard setting panels are convened for each Step, and are properly trained to provide item ratings using the Angoff procedure. There are about 12 panelists in each group. The panelists provide initial Angoff ratings on a representative subset of exam items (about 200 items) and are then given the item statistics as well as an opportunity to revise their ratings. (One of the research papers provided supported the utility of providing the panelists with the item statistics.) After providing their final item ratings, the panelists also provide Beuk/Hofstee ratings. The final standards are selected using (a) the results from two independent Angoff panels, (b) an analysis of examinee performance trends, and (c) an analysis of survey data gathered from various constituencies.

Four criterion-related validity studies were conducted to evaluate the validity of the passing standards. The passing standards have been revisited twice for each Step, and have been raised one to three points for each Step since 1993.

Although there were no passing score reliability (decision consistency) studies provided, the conditional standard error of measurement appears to be lowest at or near the passing scores.

Rating: Based on evaluation of the above, the appropriateness of the USMLE standard setting procedures was rated "exemplary."

USMLE: Evidence Supporting the Validity of Score-Based Inferences
The NBME submitted a list of 42 studies pertaining to the validity of the USMLE. Two of these studies were not considered germane to the validity of USMLE scores (one was a methodological DIF study using simulated data, the other was an MCAT validity study that used the USMLE as a criterion). The other 40 studies comprised a mix of published articles, papers presented at national and international conferences, technical reports, and technical memoranda. Several of these studies were described above (i.e., content validity or standard setting studies). Although the other studies could generally be labeled under the rubric of construct validity, they were classified into the following categories: criterion-related (20 studies), differential item functioning (3), examinee surveys (4).

Rating: Based on evaluation of the above, the evidence supporting the validity of USMLE score-based inferences was rated "exemplary."

USMLE Summary: The evidence in support of the USMLE validity argument is compelling. The design of the test specifications is impressive and well documented. Item writers are well trained. The standard setting process is exemplary. Numerous validity studies have been conducted.

Conclusion: There is a strong body of evidence that supports the USMLE validity argument.

COMLEX-USA Results

COMLEX-USA: Appropriateness of Test Development Procedures

Development of test specifications
The development of COMLEX-USA test specifications is not documented. NBOME staff reported that data from the NAMCS is used to derive test specifications; however, it is not clear how this is done. The test specifications were sent to external reviewers to gather relevance ratings, but there is no report of the results of these ratings. One limitation of the rating task is that there was no "irrelevant" category provided to the reviewers. External reviewers were asked to rate each content area as "essential," "important," or "peripheral."

Development of test items
The NBOME Development Handbook describes the item and test development processes. Faculty at osteopathic medical schools are recruited to write and review test items. Separate item writer's manuals were developed by the NBOME for each level of the exam. These manuals appear useful for training item writers to construct items of high quality. Items and exams undergo several sets of internal and external review. For 1997, 115 D.O. faculty were involved in the writing or review of COMLEX-USA items. The item and test review steps outlined in the Handbook are comprehensive and should facilitate quality tests. No reports were provided that documented the data gathered from these reviews.

Development of examinations
COMLEX-USA items are not pilot-tested, but appropriate item analysis procedures are used to validate the item keys and remove poorly functioning items before scoring. Both classical and IRT item statistics are used to evaluate items. Classical item analyses include computation of p-values, conditional p-values, point biserial and biserial correlations, "high-low" statistics, and complete distractor analyses. IRT item analyses include item infit and outfit statistics. The average item biserials are adequate.

Other content-related evidence
There is extensive documentation for candidates and lay audiences describing the test content, administration, and scoring. The Handbook indicates that content representation studies are conducted, but the results of these studies were not described or reported. No information regarding sensitivity review was provided.

Scaling and scoring examinations
All COMLEX-USA items are scored correct or incorrect. Each Level is scaled using the one-parameter (Rasch) IRT model. The fit of this model to the data is evaluated for each exam administration. Classical internal consistency reliability estimates are computed and are consistently high (above .90). Information regarding conditional standard errors of measurement was not provided, but such information is readily obtainable from the Rasch model.

Rating: Based on evaluation of the above, the appropriateness of the COMLEX-USA test development procedures was rated "satisfactory."

COMLEX-USA: Appropriateness of Standard Setting Procedures
The documentation regarding the COMLEX-USA standard setting process is sparse. Agendas from recent standard setting studies and lists of participants were provided, and a summary memo for a Level 3 standard setting procedure was provided. No information was provided regarding racial/ethnic representation of standard setting panelists. It was unclear whether any women served on the standard setting panels. Medical specialty area and geographic representation appear to be criteria used in selecting standard setting panelists.

An Angoff method is used to derive suggested passing scores. There appears to be a single standard setting panel for each Level, comprising between 9 and 16 panelists. The NBOME stated orally that a Hofstee adjustment is also used to set the standards on all three Levels. Due to the lack of documentation for Levels 1 and 2, the exact process for selecting the final passing scores on these exams is unclear.

One concern raised in the review of COMLEX-USA standard setting is that the Angoff panelists may be reviewing too many items in the amount of time allocated (800 items rated in ten hours, about 45 seconds per item). Furthermore, there appear to be no studies that evaluate the validity of the passing standard, nor studies of passing score reliability (decision consistency). Information regarding conditional standard error of measurement at the passing score was not provided.

Rating: Based on evaluation of the above, the appropriateness of the COMLEX-USA standard setting procedures was rated “insufficient.”

COMLEX-USA: Evidence Supporting the Validity of Score-Based Inferences

No validity studies have been conducted on COMLEX-USA data. Candidate surveys were conducted, which provided mostly favorable feedback regarding examination quality.

Rating: Based on evaluation of the above, the evidence supporting the validity of COMLEX-USA score-based inferences was rated “insufficient.”

COMLEX-USA Summary: Great care appears to go into the construction of the COMLEX-USA. The involvement of external D.O.s is exemplary. However, the testing program suffers from a lack of technical documentation. There is not enough evidence to support the defensibility of the standard setting procedure, and no studies have been conducted to verify that the scores demonstrate construct validity.

CONCLUSION: There is insufficient evidence in support of the validity argument for the COMLEX-USA.

Limitations

The conclusions reached in this study were based on analysis of the information provided to me by the NBME and NBOME, as well as the results of the search of electronic databases. This analysis was non-experimental. Other studies based on experimental designs, such as those discussed at the December 2, 1998 meeting should provide further information regarding the psychometric quality of the COMLEX-USA and USMLE.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, 10, 61-82.

Kane, M. (1994). Validating interpretive arguments for licensure and certification examinations. *Evaluation & The Health Professions*, 17, 133-159.

uehn, P. A., Stallings, W. M., 7 Holland, C. L. (1990). Court-defined job analysis requirements for validation of teacher certification tests. *Educational Measurement: Issues and Practice*, 21-24.

LaDuca, A., Taylor, D. D., Hill, I. K. (1984). The design of a new physician licensure examination. Evaluation & The Health Professions, 7, 115-140.National Board of Medical Examiners (1998, January). Step 3 Design and Practice Model. Philadelphia, PA: Author.

Smith, I. L., & Hambleton, R. K. (1990, winter). Content validity studies of licensure examinations. *Educational Measurement: Issues and Practice*, 7-10.

Willets-Bloom, M. C., LaDuca, A., & Henzel, T. R. (1993, April). Constructing a practice model for a physician licensing examination. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Review of Additional Standard Setting Documentation and Validity Studies Received from NBOME

Report Submitted to the Federation of State Medical Boards

Special Committee on Licensing Examinations

Stephen G. Sireci, Ph.D.
Independent Psychometric Consultant

Background

In February, 1999 I submitted a report to the Federation of State Medical Boards' (FSMB) Special Committee to Evaluate Licensure Examinations that summarized my evaluation of the psychometric evidence in support of the United States Medical Licensing Examination (USMLE) and the Comprehensive Osteopathic Medical Licensure Examination (COMLEX-USA). In that report, I concluded: (a) there was a strong body of evidence to support the argument that the USMLE is valid for the purpose of licensing physicians, and (b) there was insufficient evidence to support the argument that the COMLEX-USA is valid for the purpose of licensing osteopathic physicians.

My conclusions were based on: (a) the large number of validity studies conducted on the USMLE, (b) the comprehensive standard setting procedures used to set passing scores on the USMLE (and the documentation of these procedures), (c) the lack of validity studies conducted on the COMLEX-USA, and (d) the lack of documentation of the standard setting procedures used to set passing scores on the COMLEX-USA.

Since that time, the National Board of Osteopathic Medical Examiners (NBOME) provided new materials in support of the validity of the COMLEX-USA. Specifically, they delivered three reports documenting

their standard setting procedures (one for each of the three levels of the exam) and nineteen studies related to the COMLEX-USA or its predecessor.

In this report, I summarize my review of these new materials and discuss their impact on my evaluation of the validity argument for the COMLEX-USA.

Evaluation Criteria

My previous evaluation of the evidence in support of the COMLEX-USA and USMLE was organized using three general evaluation criteria: (a) appropriateness of test development procedures, (b) appropriateness of standard setting procedures, and (c) evidence supporting the validity of score-based inferences. For each category, each exam was given a rating of insufficient, marginal, satisfactory, or exemplary. The USMLE received ratings of "satisfactory," "exemplary," and "exemplary" for these three categories, respectively. The COMLEX-USA received ratings of "satisfactory," "insufficient," and "insufficient" for the three categories, respectively.

The newer materials provided to me by the NBOME were related to the two evaluation categories in which the COMLEX-USA was rated unfavorably: standard setting and validity studies. Based on my review of these new materials, my rating for "appropriateness of standard setting procedures" is changed from "insufficient" to "marginal," and my rating of "evidence supporting the validity of score-based inferences" is changed from "insufficient" to "exemplary." The remainder of this report provides description and rationale for these conclusions.

Description of New Materials Submitted by the NBOME

I received 22 documents from the NBOME during the last week of November 1999. Three of these documents described how the passing standards were set on each of the three Levels of the COMLEX-USA. The other 19 documents pertained to the validity of at least one of the levels of the COMLEX-USA or another exam developed by the NBOME. Twelve of these 19 studies pertained to the validity of the COMLEX-USA and 11 of these studies provided positive evidence in support of the COMLEX-USA3[3]. Six of the other documents pertained to the COMLEX-USA predecessor exams; the seventh pertained to exams developed for the American College of Osteopathic Surgeons.

Evaluation of Standard Setting Documentation

3[3] The other study (Shen & Yen, 1997) provided some evidence that the item sets (item clusters) on the COMLEX-USA promote local item dependence. Local item dependence is a rather technical issue. It is not desirable and is contrary to the scaling model underlying the COMLEX-USA and the USMLE. However, the presence of local item dependence is not a fatal flaw. I do not know of any studies on this issue pertaining to the USMLE. This study was not taken as evidence in support of the COMLEX-USA validity argument.

3[4] I did notice that one of the standard setting panelists for COMLEX-USA Level 1 also participated on the COMLEX-USA Exam Development Committee, which contradicted a statement in the report that the panelists were "independent from the Level 1 item review, test construction, and test review processes" (p. 1). This contradiction did not occur in the standard setting studies for COMLEX-USA Levels 2 and 3.

3[5]One strategy used by Schulz involved creating "category" scores by collapsing over numerous items (sometimes hundreds of items). This process obscures detection of dimensionality. A better procedure would be to create more than one parcel per category. In addition, the item-level nonlinear factor analysis results were not interpreted with respect to the test plan categories to evaluate the construct representation of the test.

Three separate reports were provided, one for each Level of the COMLEX-USA. Each report described the selection of standard setting panelists, the training of the panelists, the standard setting methods used, and how panelists' data were used to establish passing scores on each level of the COMLEX-USA. All three levels used two standard setting methods: the Angoff method and the Hofstee method. A very brief description of these methods follows.

Brief Description of Angoff and Hofstee Methods

The Angoff method begins with a discussion of the candidate who is minimally qualified to pass the exam. Panelists then review each test item and estimate the probability this minimally qualified candidate will answer the item correctly. A passing score is derived for each panelist by summing these item probabilities. The final passing score is calculated by taking the average passing score across panelists.

The Hofstee method does not require individual item judgments. Instead, panelists are asked to give their impressions of what the minimum and maximum failure rates should be for the exam, as well as what the minimum and maximum percent correct scores should be. These minimum and maximum failure rates and percent correct scores are averaged across panelists and projected onto the actual score distribution to derive a passing score. The Hofstee method is often used to evaluate or adjust the passing score derived using the Angoff method. When the two methods produce similar passing scores, they validate each other. When they diverge, executive committees typically use both sources of information to decide where to set the passing score.

The Angoff method is the most popular method used to set passing scores on licensure and certification exams (Kane, 1994; Mehrens, 1995; Plake, 1995; Pitoniak & Sireci, 1999; Sireci & Biskin, 1992). Using Hofstee judgments to evaluate or adjust the Angoff result is commendable. The National Board of Medical Examiners (NBME) also uses the Angoff and Hofstee methods to set standards on the USMLE.

Standard Setting on COMLEX-USA Level 1

Fifteen carefully selected panelists participated in the standard setting study for COMLEX-USA Level 1. Five of the panelists were DOs, 2 were MDs, 7 were PhDs, and one was a PharmD. These panelists provided Angoff ratings for all of the operational items on the June 1998 Exam, as well as Hofstee ratings for the exam.

The NBOME's selection criteria for panelists seemed appropriate, as did the training of panelists for conducting the Angoff ratings. The time allotted for panelists to complete the Angoff ratings appeared sufficient. The panelists did not take a sample of the items, without the answer keys, to get a realistic idea of the difficulty of the items. Other than that omission, the Angoff study appeared to be well executed.

The passing score derived using the Angoff method resulted in a significantly lower failure rate than the passing score derived using the Hofstee method. The report provided justification for using the higher passing score based on the Hofstee method.

In general, the process used to set the passing score on the COMLEX-USA appears to be reasonable and defensible4[4]. In addition, the report provided information on the standard error of measurement associated with the passing score. Relative to other points along the exam score scale, the error of measurement was at a minimum, which is desirable.

Standard Setting on COMLEX-USA Level 2

The standard setting procedures used for COMLEX-USA Level 2 were similar to those described above for Level 1. Twelve carefully selected panelists participated in the study (11 DOs and one EdD). The training of the panelists and the time allotted for gathering their ratings seemed appropriate. Once again, the passing standard derived using the Angoff method was noticeably lower than the passing standard derived using the Hofstee method. The Executive Committee decided to use the higher (more rigorous) passing standard. The rationale provided for this decision seems reasonable. In general, the process used to set the passing score on the COMLEX-USA Level 2 appears to be reasonable and defensible. In addition, the error of measurement was lowest at the passing score, which is desirable.

Standard Setting on COMLEX-USA Level 3

Nine carefully selected panelists participated in the standard setting study for COMLEX-USA Level 3 (all 9 were DOs). The following (excerpted) summary description of the candidate who is minimally qualified to pass the exam was provided in the report:

borderline examinees tend to be "single-minded," or easily confused or overwhelmed when more clinical information is presented…they do not have the ability to quickly synthesize information. Borderline examinees usually are slow to reach patient management decisions. They might know all of the options and all of the consequences of each of the options, but the decisions they make often address the trivial aspects but miss the major point. (p. 2-3)

As described earlier, the notion of the minimally qualified candidate is critical to the Angoff method. I do not know the minimal knowledge, skills, and abilities candidates who pass COMLEX-USA Level 3 should have, but this description struck me as potentially lower than it should be. I did not get that impression from the descriptions provided in the documentation for COMLEX-USA Levels 1 or 2. Also, the standard setting documentation I reviewed for the USMLE did not include descriptions of the minimally qualified candidates, so I cannot use the USMLE descriptions as a basis for comparison.

The training of the panelists and the time allotted for gathering the Angoff ratings seemed appropriate. The passing standard derived using the Angoff method was considered to be too low. An adjusted Angoff passing score was calculated by eliminating those panelists whose ratings appeared to be inconsistent with the others, and those items that seemed out of line with the others. This adjusted passing standard was higher than the passing standard derived using the unadjusted Angoff data, but was still considerably lower than the passing score interval suggested by the Hofstee data. Furthermore, the Hofstee data could not be used to derive a passing score because panelists' ratings were inconsistent with candidates' performance on the exam (this finding happens occasionally with the Hofstee method, which is one of its weaknesses). The NBOME staff suggested a passing score that fell above the adjusted Angoff passing score and below the lower bound of the passing score interval suggested by the Hofstee data to the NBOME Executive Committee. The Committee accepted this recommendation. The rationale for this decision was not as convincing as the rationale provided for COMLEX-USA Levels 1 and 2. In particular, it described the adjusted Angoff method as being a "more valid index" than the unadjusted Angoff passing standard; however, there was no evidence to support that statement. Similar to Levels 1 and 2, the conditional standard error of measurement was lowest at the passing score, which is desirable.

In general, the process used to set the passing score on the COMLEX-USA Level 3 seems marginal. The NBOME attempted to implement a proper standard setting study, but several problems occurred. First, three panelists selected to participate in the meeting were unable to attend. The use of only nine panelists to set a passing standard is less than desirable. The elimination of an unspecified number of panelists to derive the adjusted Angoff passing score means even fewer panelists were used to derive this passing standard. Second, the results suggest it is possible that the Angoff data could be unreliable and the panelists' Hofstee ratings were unrealistic. These two observations strongly argue for a reevaluation of the COMLEX-USA Level 3 passing standard.

It should be noted that several documents state the COMLEX-USA passing standards are reexamined every three years. For example, Osborn, Meoli, Buser, Clearfield, Bruno, and Sumner-Traux (1999) state "NBOME policy requires reexamination of Pass/Fail standards for each level of the COMLEX-USA every three years" (p. 14). The standard setting study for Level 3 of the COMLEX-USA was conducted in February 1995. Although the passing standard may have been reevaluated from a policy perspective, it appears no other standard setting study has been done on this exam since 1995.

Conclusion Regarding COMLEX-USA Standard Setting

By providing reports on the COMLEX-USA standard setting processes, the NBOME fulfilled its obligations (as stipulated by professional standards for testing) to document the process and provide this information to interested parties. The procedures used to set the passing scores on COMLEX-USA Levels 1 and 2 appear to be reasonable and defensible. The procedure used to set the passing score on COMLEX-USA Level 3 is less defensible. Although I am sure it would hold up in court, the results were sufficiently anomalous to warrant a new study. It is important to note that in all three standard setting studies, the adjustments to the Angoff standards were in the direction of a more rigorous standard, which is commendable from the perspective of protecting the public. Future NBOME research should focus on why the Angoff standards are so low. The relatively large and consistent differences across the Angoff and Hofstee methods warrant further study.

Level 3 of the COMLEX-USA is a critical component in the licensure process for osteopathic physicians. Given the questions regarding the passing score for COMLEX-USA Level 3, and the lack of a significant follow-up study for that standard over the past five years, I rate the appropriateness of the standard setting procedures for the COMLEX-USA as "marginal."

Comparing the COMLEX-USA and USMLE Standard Setting Processes

In my evaluation of the standard setting procedures for the USMLE, I rated the appropriateness of these standard setting procedures as "exemplary." A few words describing why the USMLE procedures were rated "exemplary" and the COMLEX-USA procedures were rated "marginal" are warranted.

An impressive feature of the USMLE standard setting procedure is that they use two independent standard setting panels to set the standard for each exam. Each panel comprises about 12 panelists. The use of independent panels allows for evaluation of the consistency of the passing standard across panels and provides twice as much information for the Executive Committee to decide where the passing standard should be set. In those cases where the results of one panel are questionable, as was the case for the COMLEX-USA Level 3, the results from the other panel can be given more weight. In addition, the passing standards for each exam have been reevaluated twice since their implementation in 1993. Finally, in selecting standard setting panelists, inclusion of minorities was a selection criterion. The composition of the standard setting panels for the USMLE was much more diverse with respect to sex and ethnicity than the COMLEX-USA standard setting panels.

Evaluation of COMLEX-USA Validity Studies

As mentioned earlier, I received 19 documents related to validity evidence for the COMLEX-USA. Seven of these documents reported research on other exams. Twelve studies pertained to the COMLEX-USA and eleven provided validity evidence in support of the COMLEX-USA. The 19 studies received are summarized in Table 1, which lists a citation for each study, whether it could be construed as a COMLEX-USA validity study, the type of validity evidence provided, and a rating regarding the strength of the validity evidence provided.

As indicated in Table 1, I judged four studies as providing substantial validity evidence in support of the COMLEX-USA validity argument. Three of the studies focused on Level 1, the other study focused on all three Levels. These four studies provided "criterion-related" evidence of validity, which means they

investigated the relationships among COMLEX-USA scores and other relevant criteria. The validation criteria investigated in these studies included grades in osteopathic medical school (all four studies), school administrators' classifications of students (Baker, Foster, et al., 1999), and pre-admission criteria (Baker, Cope, et al., 1999). Three of the studies were school specific studies (two at the West Virginia School of Osteopathic Medicine, one at the Oklahoma State University College of Osteopathic Medicine), and one was a multi-school study (Baker, Foster, et al., 1999). The Baker, Foster, et al. study was particularly impressive in that it involved almost the entire population of a cohort of osteopathic students (18 of the 19 osteopathic medical schools participated in this study). All four studies provided strong evidence that COMLEX-USA scores were positively related to students' grades in medical school (in general correlations ranged from .65 to .80, r-squares were typically .5 or higher). The relationships among COMLEX-USA scores and pre-admission criteria were very low, but that finding was consistent with the researchers' expectations, given the admission criteria used at that particular school. The relationship between COMLEX scores and administrators' ratings of students was also positive. These studies all provided strong criterion-related evidence in support of the validity of COMLEX-USA scores for the purpose of licensing osteopathic physicians. The degree of evidence provided is on par with similar studies conducted on the USMLE.

Shen and Yen (1999, August) investigated the relationship between COMLEX-USA Level 3 scores and postgraduate NBOME specialty exams. Strong relationships were found among COMLEX scores and four separate specialty exams (minimum r-squares were around .4). Presently, this study is an abstract that does not give information regarding sample sizes, where the data came from, and other aspects of the study. When the final report is written, the information contributed by this study could be upgraded from "moderate" to "substantial."

Shen and Yen (1999, April) investigated the relationship among students' performance on COMLEX-USA Levels 1 and 2 and NBOME subject exams. This report, which is also an abstract, reported correlations ranging from .55 to .87. The magnitudes of the correlations were consistent with the content of the different subject exams. This study provided concurrent and predictive criterion-related evidence in support of the validity of COMLEX-USA scores.

Shen (1999, March) investigated group differences and differential item functioning across students from organ system-based and discipline-based osteopathic medical schools. The study focused on a subset of 70 items from COMLEX-USA Level 1. Differences across these two educational orientations were found, but they were extremely small. There was no systematic item bias against either group of students. This study was an interesting exploration of the construct validity of COMLEX-USA Level 1 scores. The results support the validity of scores from this exam.

Osborn et al. (1999) provided an overview of the logic underlying the development of the COMLEX-USA as well as a comprehensive description of the content of all three Levels of the exam. It is recommended reading for Committee members who want to become more familiar with the content of the COMLEX-USA. This report did not involve the analysis of any COMLEX-USA data, but it does provide evidence in support of the content validity of the exam (evidence in support of domain definition). I found one technical flaw in this manuscript. It states the validity question for the COMLEX-USA is "Does the examination measure what we say it's measuring?" A more appropriate and critical validity question is "Are COMLEX-USA scores useful for deciding who is competent to practice as an osteopathic physician?"

Schulz (1999) investigated the dimensionality of the COMLEX-USA. This investigation is important for at least three reasons. First, the theoretical structure of the COMLEX-USA specifies two dimensions underlying all three levels. Corroboration of the existence of these dimensions by empirical analysis would support the construct validity of the exam scores. Second, the scaling and scoring models for the COMLEX-USA assume the items on each level can be adequately scaled using a single (dominant) dimension. The viability of a single dimension underlying the data needs to be evaluated. Third, recent manuscripts (e.g., Shen, 1999; Wright & Bezruckzo, 1999) claim the same dominant dimension underlies all three levels of the COMLEX-USA and so osteopathic students' growth can be measured across the three exams.

The Schulz (1999) report provided some evidence that each Level of the COMLEX-USA can be scaled (separately) using a single dimension. The manuscript also reported that approximately 100 items are common across the three Levels of the COMLEX-USA (p. 3). The results of the study provided some evidence that these "anchor" items are measuring the same dimension as the other items on each level. The study did not provide evidence that all three Levels of the COMLEX-USA are measuring the same dimension. Therefore, claims that all three Levels can be equated and used to measure growth remain unsubstantiated. Also, the results did not reflect the hypothesized structure of the exam, but that finding could be a result of the particular methodologies employed5[5]. Finally, it should be noted that in analyzing data for COMLEX-USA Level 1, Schulz used data from the second administration of the exam, which involved only 183 examinees. Since second-time takers represent a different population than the first administration of the exam, it is entirely possible the results of this study would not generalize to the higher volume administration of Level 1. This sample size is also very small for the purposes of assessing dimensionality.

Shen (1999, April) investigated the statistical performance of 15 osteopathic principles and procedures items across different osteopathic medical schools. The study was interesting from both methodological and construct validity perspectives. Four of the 15 items 27%) functioned differently across schools. Content analyses of these items are planned to inform future COMLEX-USA test development activities.

Shen (1999, November) investigated whether the three Levels of the COMLEX-USA can be used to measure osteopathic candidates' growth over the course of their training. The study looked at students' performance on 100 items that were common across all three Levels (presumably, the same items discussed in Schulz, 1999). The items were more difficult for students taking Level 1 than for students taking Levels 2 or 3. However, the students taking Level 1 came from the second administration of this exam and cannot be considered representative of all students taking Level 1. This study has only been partially completed. The present version does not provide much validity evidence in support of the COMLEX-USA. The study claimed "The COMLEX design raises the standard of the quality for the [sic] medical licensure in that it addresses the issue of continuity and differentiation of various components of medical licensure examination" (p. 1). No evidence was provided to support that claim.

The other studies listed in Table 1 do not provide evidence in support of the COMLEX-USA validity argument. Shen and Yen (1997) was mentioned earlier (see footnote 1). Shen, Cavalieri, et al. (1997) investigated whether two groups of osteopathic residents differed with respect to their skills, as measured by the COMLEX-USA Level 3. It was not framed as a validity study and does not provide validity evidence in support of the exam. The other studies did not involve analysis of COMLEX-USA data. Five of the studies involved concurrent calibration of all three parts of the previous NBOME licensing exams, but the adequacy of the equating is questionable. The last study (Shen, 1999, in press) is an interesting study regarding measuring change in osteopathic residents' knowledge, but it did not involve analysis of COMLEX-USA scores.

Summary of COMLEX-USA Validity Studies

The NBOME and researchers at two osteopathic medical schools conducted recent studies that provide important evidence in support of the validity of COMLEX-USA scores. Seven studies provided criterion-related evidence in support of using the COMLEX-USA to license osteopathic physicians. Other studies are underway to explore the structure of this exam and evaluate its utility for measuring osteopathic students' growth in medical knowledge over time. Based on these studies, I rate the evidence supporting the validity of score-based inferences for the COMLEX-USA as "exemplary."

Conclusion

Based on: (a) my previous review of technical documentation from the NBOME, (b) my analysis of the standard setting documentation for the COMLEX-USA, and (c) my review of the recent validity studies on this exam, I conclude there is sufficient evidence to support the argument that the COMLEX-USA is valid for the purpose of licensing osteopathic physicians.

References

Baker, H. H., Foster, R. W., Bates, B. P., Cope, M. K., McWilliams, T. E. Musser, A., & Yens, D. (1999, November). Relationship between achievement in the first two years of osteopathic medical school with performance on the June 1999 COMLEX Level 1 Examination. (Unpublished manuscript).

Baker, H. H., Cope, M. K., Foster, R. W., Gorby, J. N., & Fisk, R. (1999, October). Relationship of academic performance during the fist two years of the curriculum with performance on the June, 1999 COMLEX-USA Level 1 Examination. (Unpublished paper).

Baker, H. H., Cope, M. K., Fisk, R., Gorby, J. N., & Foster, R. W. (1999, June). Relationship of pre-admission variables and first- and second-year course performance to performance on the National Board of Osteopathic Medical Examiners' COMLEX-USA Level 1 Examination. (Unpublished manuscript).

Kane, M. (1994). Validating interpretive arguments for licensure and certification examinations. *Evaluation & The Health Professions*, 17, 133-159.

Lapolla, M., Goodson, L. B., & Pickard, T. (1999). The COMLEX testing process: An institutional Research White Paper. Tulsa, OK: Center for Health Policy Research, Oklahoma State University.

Mehrens, W. A. (1995). Methodological issues in standard setting for educational exams. Proceedings of the *Joint Conference on Standard Setting for Large Scale Assessments of the National Assessment Governing Board* (NAGB) and the *National Center for Educational Statistics* (NCES), Volume II (pp. 221–263). Washington, DC: U. S. Government Printing Office.

Osborn, G. G., Meoli, F. G., Buser, B., Clearfield, M. B. Bruno, J., & Sumner-Truax, L. (1999, November). The Comprehensive Osteopathic Medical Licensing Exam, COMLEX-USA: A new paradigm in testing and evaluation.

Pitoniak, M. J., & Sireci, S. G. (1999). A literature review of contemporary standard-setting methods appropriate for computerized-adaptive tests. Laboratory of Psychometric and Evaluative Research Report No. 369. School of Education, University of Massachusetts, Amherst, MA.

Plake, B. S. (1998). Setting performance standards for professional licensure and certification. *Applied Measurement in Education*, 11 (1), 65–80.

Schulz, E. M. (1999, September). An assessment of the dimensionality of the Comprehensive Osteopathic Licensing Examination (COMLEX-USA). (Unpublished manuscript).

Shen, L. (1993, April). Constructing a measure for longitudinal medical achievement studies by the Rasch model one-step equating. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Shen, L. (1994). Gender effects on student performances on the NBOME Part I, Part II, and Part III. Academic Medicine, 69 (supplement), S75-S77.

Shen, L. (1994, April). The growth patterns of general medical achievement. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Shen, L. (1995, June). Factors influencing the growth of general medical knowledge of osteopathic medical students. NBOME Research Report.

Shen, L. (1995). The growth patterns of general medical achievement. Proceedings of the *Sixth Ottawa International Conference on Medical Education*.

Shen, L. (1999, March). Curriculum type and content fairness of the Comprehensive Osteopathic Medical Licensing Examinations (COMLEX). NBOME Research Report.

Shen, L. (1999, April). A multilevel assessment of differential item functioning in the COMLEX due to the diversity of school curriculum. Paper presented at the annual conference of the American Educational Research Association, Montreal, Canada.

Shen, L. (1999, August). Performance on COMLEX Level 3 and osteopathic postgraduate specialty written tests. NBOME Research Abstract.

Shen, L. (1999, November). COMLEX: A progressive medical licensing examination series. NBOME Research Report.

Shen, L. (1999, in press). Progress testing for postgraduate medical education: A four-year experiment of American College of Osteopathic surgeons resident examinations. Advances in Health Sciences Education.

Shen, L., Cavalieri, T., Clearfield, M., & Smoley, J. (1997). Comparing medical knowledge of osteopathic medical trainees in DO and MD programs: A random effect meta-analysis. *Journal of the American Osteopathic Association*, 97, 359-362.

Shen, L., & Yen J. (1997). Item dependency in medical licensing examinations. *Academic Medicine*, 72 (supplement), S19-S21.

Shen, L., & Yen J. (1999, April). Relationship between student performances on COMLEX-USA and NBOME subject exams. NBOME Research Abstract.

Sireci, S. G., & Biskin, B. J. (1992). Measurement practices in national licensing examination programs: A survey. *CLEAR Exam Review*, 21–25.

Wright, B. D., & Bezruczko, N. (1999, July). Evaluation of COMLEX-USA. Chicago, IL: MESA Psychometric Laboratory, Statistical Laboratory, Department of Education, University of Chicago.

Attachment 4

Overview of Comparability Studies

The first study was to review candidates who have taken both the COMLEX-USA and the USMLE in order to evaluate this cohort's passing percentages on each test, thereby potentially providing information on the relative difficulty differences between the examinations. The study would have used a sample of approximately 300 students who have taken both COMLEX-USA Level 1 and USMLE Step 1, and as such, would not be a representative sample, as no MDs have taken the COMLEX-USA. Initially, the committee

determined that such a study would provide valuable information and requested this study be conducted in the absence of the ability to commission alternate, more representative studies. However, upon reevaluation the committee determined that, although the study would be inexpensive and easily accomplished, the results would provide limited information and the study itself had too many confounds. Therefore, the committee agreed to reassess the feasibility of conducting the study pending review of other options.

The second study reviewed by the committee was a content validity study to (1) appraise the content quality of COMLEX-USA and USMLE items, (2) appraise the relevance of COMLEX-USA and USMLE items with respect to the practice of general medicine and (3) discover similarities and differences among COMLEX-USA and USMLE items with respect to the knowledge and skills they measure. At the request of the Committee, the Federation's Board of Directors approved funding to conduct the study contingent upon the participation of the NBOME and the NBME in providing the necessary information. Although the approved budget did not reflect costs to the NBME and the NBOME and no provision for reimbursement to either organization was provided, both organizations indicated their willingness to participate in the study assuming the costs were manageable. Dr. Sireci met with representatives from both organizations to discuss the logistics of implementing the study, and both organizations made a commitment to see the study through completion. However, in order for the study to be valid, the number of items reviewed from each Step/Level of each examination and the number of expert reviewers would need to be expanded considerably, which would result in a substantial increase in cost. Even after increasing the number of items and reviewers, there were still problems to be addressed such as how to select DO and MD reviewers in such a way as to ensure parallel representation of practice areas within each group, and how to select items that come from multiple-item scenarios. Additionally, the study would not provide information regarding the comparability of the USMLE and the COMLEX-USA's content quality or rigor and would only reinforce evidence regarding the comparability of the content and structure of the two examinations already gained from previous studies. Furthermore, the committee determined that the information gained would not aid it in developing a recommendation about the Federation's policy regarding a single pathway to licensure. Therefore, the committee decided not to conduct the content validity study.

The final study discussed by the committee was designed to look at the similarity of the construct(s) measured, requiring the creation of one or more "mixed" test forms that would comprise questions from both the COMLEX-USA and USMLE. The experimental forms would be administered to a sample of 200 or more medical students. Dimensionality analyses would be conducted to see if items from the two exams load on common dimensions. The empirical difficulties of the two sets of items could also be evaluated. If both DO and MD candidates were tested, comparisons among the groups could be made as well as "predicted" pass/fail differences on both tests. Correlations between subscores on each "test" could be calculated, and these subscores could be correlated with medical school grades to look at differential predictive validity. The Federation Board of Directors deferred action on the committee's request to explore funding sources. Subsequently, the committee determined the study to be cost prohibitive and agreed to reassess the need for the construct validity study after evaluating the results of the content validity study.