ABSTRACT
Developing test content for the USMLE involves significant efforts from physician volunteers and staff associated with the program. The bedrock of this process takes place among the test materials development committees (TMDCs) where physicians and content experts write multiple-choice questions for all three USMLE Steps. Ongoing assessment of the item pool by the respective Step Committees initiates item-writing assignments that bolster or maintain content in specific areas. Staff at the National Board of Medical Examiners (NBME) then assist item-writers to assure a consistent style and structure for all USMLE test items. All test materials are crafted to complement an overall examination blueprint. Multiple levels of review and pre-testing ensure that all test items making their way onto examination forms as live or ‘scored’ material are appropriate, statistically sound and presented in test forms balanced to be consistent with the content outline and examination blueprint.

INTRODUCTION
Since its implementation in 1992, the United States Medical Licensing Examination® (USMLE) has provided state medical boards with a high quality, standardized national tool for assessing physician knowledge prior to issuing an initial license for unsupervised medical practice. Today, all allopathic and composite medical boards require successful completion of the USMLE as a condition for licensing their M.D. degreed physician candidates.1

This article continues the periodic series on the USMLE begun in 2005. Prior articles in the series focused on a broad introductory overview of the program, the Step 2 Clinical Skills (CS) examination and the program’s processes for maintaining examination security.2,3,4 The intent of this article is to provide readers with an understanding of the committee structure and processes for developing USMLE examination content with a particular focus on the development of multiple-choice questions for the exam.

COMMITTEE STRUCTURE
While the USMLE is a joint program of the Federation of State Medical Boards (FSMB) and the National Board of Medical Examiners (NBME), the development of examination content is really a collaborative effort involving the talents and efforts of many individuals working beyond the walls of these two organizations. Much of the work in writing and reviewing test materials is performed by physicians and clinicians drawn from across the country and representing multiple perspectives: the medical licensing community, academic medicine and clinical practice. In this sense, the USMLE relies upon a “national faculty” of experts numbering more than 300 strong and serving on approximately 40 committees.5,6

Program governance is conducted through the USMLE Composite Committee, whose appointed members represent the FSMB, the NBME, the Educational Commission for Foreign Medical Graduates (ECFMG) and the American public. The Composite Committee is charged with broad responsibilities for the program, e.g., approval of the exam blueprint for each Step, establishing program policy as well as scoring and standard setting systems.

An examination committee has been established for each of the three USMLE Steps. They are the Step 1, Step 2 and Step 3 Committees. These committees operate under the auspices of the Composite Committee and are charged with designing their respective Step’s design, determining testing methods, supervising test item development, approving test forms and setting the pass/fail standard. Appendix 1 offers a description of the USMLE process for standard setting.
Two additional levels of committee work are critical to the development of USMLE test content. Supporting each Step committee are several Interdisciplinary Review Committees (IRCs) and multiple Test Materials Development Committees (TMDCs) (see Figure 1). The latter serves as the foundational base for USMLE test development as it is the members of these committees who write the test items that ultimately appear on the USMLE. The majority of the USMLE program’s national faculty work as item writers on TMDCs. The efforts required by the TMDC members to produce high quality test items far exceed the modest rewards offered in return: a small honorarium and limited number of continuing medical education hours. The IRCs provide a review and quality control function for materials developed by the TMDCs. This is covered more fully in the section “The IRC: Review and Approval of Live Materials.”

Newly appointed TMDC members attend a multi-day item-writing workshop in Philadelphia conducted by NBME staff to orient the new members to the mechanics and style of writing test questions for the USMLE. Afterward TMDC members receive assignments to write questions in their area of expertise. When a TMDC committee re-convenes in Philadelphia, they will collectively review and critique drafted items and review performance data for items that have been pre-tested previously with examinees. See Figure 2 for an overview of the test development process.

In speaking with current and former members of TMDCs, a common theme often arises from the conversation. For the committee member, the true rewards of participation stem from the collegial nature of the test development enterprise, the opportunity to meet and interact with fellow physicians from across the country and a satisfaction that they are “giving back” to the medical profession they love.

Developing multiple choice questions (MCQs) for USMLE
While the USMLE provides insight into clinical and communication skills through Step 2 CS and patient management through the Primus computer case simulations on Step 3, MCQs comprise the majority of the content for assessing physician knowledge in the USMLE sequence. In 1999, the USMLE program moved to a computer-based form of test administration with testing offered year-round. The latter element requires a test pool of many thousands of items for each Step. This allows the USMLE program to create multiple test forms for each Step while minimizing any duplication in content that examinees will see. Maintaining a test pool of high quality MCQs for each Step is a critical activity of the program.

Pool Analysis and Assignment of Items
Before items are assigned or written for a USMLE Step, an analysis of the item pool is conducted to identify topic areas in the pool that are shallow or deep. The purpose of this exercise is to identify areas of the item pool that need particular focus in a given year to build up test content. TMDC members are asked to write new items in shallow areas to level the test pool and increase the number of non-overlapping test forms that can be constructed.

Preparation and Submission of New Items
TMDC members are typically asked to write approximately 50 new items annually. This allows USMLE to address shallow areas or replace content scheduled for retirement. The work of the TMDC members is done at their home or home institution. Items, including any associated pictorial materials, are submitted to NBME editorial staff. The USMLE encourages item writers to include pictorial materials with a large percentage of their items. These pictorial materials may include graphs or drawings, clinical photo-
Original questions sent to NBME by author

Items edited; returned to author

Author changes incorporated

Committee meets to review all questions and approve items for pretesting

ITEM PRETESTING

After testing by sufficient # of examinees, pretested questions triaged according to pre-established statistical criteria

Pretested materials with GOOD statistics

Pretested materials with BAD statistics

Interdisciplinary review committee (IRC) meets to approve items for use in scored section of the exam (expiration date set for next review by IRC)

Exam administered

SCORED questions separated according to pre-established statistical criteria and expiration date

Expanding items with good statistics return to IRC for reapproval/revision

Items with good statistics will continue to circulate for as many years as designated by IRC (default = 3 years)

Staff intervention
Delete from pool or return the item to original committee for major rewrite, minor rewrite or deletion
graphs depicting physical findings, gross or histopathological specimens or results of commonly encountered diagnostic studies (e.g., ECGs, x-rays, MR scans). In 2007, the USMLE program began including a small number of test items in Step 2 Clinical Knowledge (CK) that use audio and/or video clips of physical findings and doctor-patient interactions. Similar multimedia items were phased into Step 1 and Step 3 in 2008.

**Interim Editing**

Upon receiving items from TMDC members, NBME test development staff place the items into a tracking table and verify that all key components required for each item have been included (i.e., answer key, content classifications, associated pictorials); authors are contacted to supply any missing information. Figure 3.1 illustrates how an item might read at this early stage of development. Most items are in the form of a patient vignette in which the first sentence provides the patient age, gender, site of care, presenting complaint and its duration. Subsequent sentences in the vignette provide additional patient history, physical findings, the results of diagnostic studies and/or response to initial treatment. A staff editor reviews the item to see that it conforms to the requested USMLE style and to ensure no information is missing. Staff also edit and annotate items for clarity, grammar and punctuation, uniformity of style and technical item flaws – particularly those that might otherwise benefit test-wise examinees or add irrelevant difficulty. Edited items (Figure 3.2), along with the original versions, are returned to item writers for revision and approval before being incorporated into a draft for review at the TMDC meeting.

**Review and Approval of Item Revisions by Item Authors**

Authors review their edited items, respond to queries from the staff editor, verify the correct answer and classification codes, and confirm the appearance of any associated pictorials. Any disagreements about phrasing are generally negotiated between the editor and author in order to arrive at a consensus about the version to be included in the draft of materials for review at the TMDC meeting. On the rare occasions when consensus cannot be reached, both the author’s and editor’s version are included in the draft.

**Review and Approval for Pre-testing at the TMDC Meetings**

A draft of test materials that includes the final approved version of each item is mailed to all TMDC members prior to their scheduling meeting. During the three-day meeting, all items are read aloud by the author, and a decision is made by the TMDC to accept, rewrite or delete each item. A staff editor facilitates discussion, assists in refining items and maintains an official record of all committee decisions, including text and classification changes and final disposition of the item. The committee chair assigns a quality grade for each item. These grades are used when selecting items for placement in examination forms. The overall acceptance for items reviewed at TMDC meetings is typically at or above 90 percent.

**Pretesting**

Following the TMDC meeting, all accepted items are updated to reflect the final phrasing approved at the meeting. The revised items receive one more review by the editor to ensure accuracy and adherence to style. An assigned staff proofreader then reviews all items for grammar, punctuation and uniformity of style. (Figure 3.3) Any questions that arise about content during this phase are discussed with the relevant TMDC chair and revised as needed.

Finalized items are uploaded into the NBME item bank and made available for pre-testing. Pretest items are included in live test forms administered to examinees; however, as pretest material they are unscored and, thus, not used in determining the examinee’s pass/fail outcome for that administration. Pretest items are not identified as unscored material to the examinee. In this way, an accurate assessment of the item’s performance can be obtained. Each MCQ is pre-tested by a minimum of 200 examinees in order to project the statistical characteristics of the item (e.g., item difficulty, discrimination†).

**The IRC: Review and Approval for Live Materials**

The purposes of the IRCs are to (1) annually review and approve newly pre-tested items (along with statistical performance) into the live, i.e., scored, pool and (2) re-review and re-approve expiring items currently in the live pool for continued use. Once approved for live use, each item is scheduled for re-review for continuing use three years later. Approximately one-third of the live (i.e., scored) pool is reviewed annually for continued accuracy and relevance.

In order for pre-tested items to be selected for review by an IRC, the item difficulty (i.e., proportion of students answering the question correctly or p-value) and discrimination index must meet prescribed statistical criteria. These vary somewhat by Step. In general, however, MCQs must have an item difficulty or p-value of greater than 30 percent and less than 97 percent. The discrimination index must have a positive correlation (i.e., most examinees from
3.1 Original Question Sent by Author
A six-year-old boy is brought to the physician by his mother because of concern about his short stature. He feels well. He has no history of serious illness and takes no medication. Immunizations are up-to-date. He is at the 7th percentile for height and the 11th percentile for weight. Examination shows a large head, a prominent forehead and mid face hypoplasia. There is lumbar lordosis. Examination of the upper and lower extremities shows decreased length, a tridentate appearance of the fingers on extension, and a bowleg deformity. Genetic testing shows a mutation of the code for fibroblast growth factor receptor three. Which of the following is the most likely diagnosis?
(A) Achondroplasia
(B) Osteogenesis imperfecta
(C) Patient has autosomal recessive disorder
(D) Patient has a sex linked disorder
(E) Most patients with this disorder also have abnormal procollagen synthesis
(F) Growth hormone has no effect on bone growth in this abnormality

3.2 Initial Edit: Returned to Author
A six-year-old boy is brought to the physician by his mother because of concern about his short stature. He feels well. He has no history of serious illness and takes no medication. Immunizations are up-to-date. He is at the __th percentile for height and the __th percentile for weight. Examination shows a large head, a prominent forehead and mid face hypoplasia. There is lumbar lordosis. Examination of the upper and lower extremities shows decreased length, a tridentate appearance of the fingers on extension, and a bowleg deformity. Genetic testing shows a mutation of the code for fibroblast growth factor receptor three. (OK to omit or replace with x-ray findings? Would a pediatrician have genetic testing done for this case?) Which of the following is the most likely diagnosis?
(A) Achondroplasia
(B) Osteogenesis imperfecta
(C) ____________________? [specify an autosomal recessive disorder]
(D) ____________________? [specify a sex linked disorder]
(E) ____________________?
(F) ____________________?

Note: Please rewrite as shown to eliminate the true/false format. Also, would this have been diagnosed earlier in this patient? Can he be younger?

3.3 TMDC Version Incorporating Author’s Changes
A six-year-old boy is brought to the physician by his mother because of concern about his short stature. He feels well. He has no history of serious illness and takes no medication. Immunizations are up-to-date. He is at the 7th percentile for height and the 11th percentile for weight. Examination shows a large head, a prominent forehead and a flat nasal bridge. There is lumbar lordosis. Examination of the upper and lower extremities shows decreased length, a tridentate appearance of the fingers on extension, and a bowleg deformity. Genetic testing shows a mutation of the code for fibroblast growth factor receptor three. Which of the following is the most likely diagnosis?
(A) Achondroplasia
(B) Osteogenesis imperfecta
(C) Gonadal dysgenesis 45, X (Turner syndrome)
(D) Leri-Weill syndrome
(E) Noonan syndrome

the top half of performers must answer the question correctly and most examinees from the lower half of performers must miss the question).

A draft of test materials is mailed to all IRC members prior to their scheduled meetings. In preparation for the multi-day meeting, members are assigned to review a specific subset of items; the reviewers are responsible for presenting these items during the meeting. For pre-tested items, reviewers are instructed to consider the appropriateness of each item
3.4 Administered in Exam as Pretest

Item Stem: SCBB5064

A six-year-old boy is brought to the physician by his mother because of concern about his short stature. He feels well. He has no history of serious illness and takes no medication. Immunizations are up-to-date. He is at the 7th percentile for height and the 11th percentile for weight. Examination shows a large head, a prominent forehead and a flat nasal bridge. There is lumbar lordosis. Examination of the upper and lower extremities shows decreased length, a tridentate appearance of the fingers on extension, and a bowleg deformity. Which of the following is the most likely diagnosis?

(A) Achondroplasia
(B) Gonadal dysgenesis 45,X (Turner syndrome)
(C) Leri-Weill syndrome
(D) Noonan syndrome
(E) Osteogenesis imperfecta

Exam Admin Exam Instance Use Bk Pres Seq p rb r Pop Medley ID
Step 2-0301 STP2C/ST9 P S 0001 8636 83 +20 +14 283 MBK1802

Subgroup Statistics: STEP 2–0301 STP2C/ST9 MBK 1802
*M* *O* A B C D E
0 0 91 0 0 5 4 High
0 0 71 0 9 13 7 Low

3.5 Administered in Exam as Live

Item Stem: SCBB5064

A six-year-old boy is brought to the physician by his mother because of concern about his short stature. He feels well. He has no history of serious illness and takes no medication. Immunizations are up-to-date. He is at the 7th percentile for height and the 11th percentile for weight. Examination shows a large head, a prominent forehead and a flat nasal bridge. There is lumbar lordosis. Examination of the upper and lower extremities shows decreased length, a tridentate appearance of the fingers on extension, and a bowleg deformity. Which of the following is the most likely diagnosis?

(A) Achondroplasia
(B) Gonadal dysgenesis 45,X (Turner syndrome)
(C) Leri-Weill syndrome
(D) Noonan syndrome
(E) Osteogenesis imperfecta

Exam Admin Exam Instance Use Bk Pres Seq p rb r Pop Medley ID
Step 2-0401 STP2C/ST10 L S 0001 5665 83 +29 +19 1078 MBK1802

Subgroup Statistics: STEP 2–0301 STP2C/ST9 MBK 1802
*M* *O* A B C D E
0 0 92 0 2 5 1 High
0 0 76 1 2 12 9 Low

for the examination purpose (e.g., all items approved for use on the Step 2 CK exam must be appropriate for all new interns regardless of specialty) and to verify all classification codes. Live items are reviewed for currency and continued appropriateness for the examination’s purpose.

At the meeting of an IRC, all items are read aloud by the assigned reviewer who makes a recommendation about disposition (See Figure 3.4 and 3.5 for examples). The committee then takes one of the following actions for disposing of each pretest and expiring live item.
• Approved for inclusion in the live, (i.e., scored) pool – The item is accepted as written.
• Return for pre-testing – Minor revisions are required that can be made at the meeting but the item will be sent back to be pre-tested again.
• Send back to TMDC for revision – Content is acceptable but major revision is required.
• Delete from the pool – Content is inappropriate.

A staff editor facilitates discussion and records all committee decisions, including classification changes and final disposition of the item, as well as any notes to be sent back to the TMDCs.

Test Form Approval

Once test items have been approved by the IRC for inclusion in the live pool, these materials are then available for placement in test forms (as scored items) for their respective Step examinations. The Step Committees are responsible for review and approval of their respective test forms. This occurs annually during a multi-day meeting when the Step Committee reviews each test form prior to its utilization in test administration. Prior to this meeting, staff has already created multiple parallel test forms following the exam blueprint previously established and approved by the respective Step Committees.

Year-round testing in a computer-based format requires thousands of test items and multiple forms of each USMLE Step examination. The focus of the Step Committees in reviewing each test form is to ensure appropriate ‘balance’ in each form (i.e., that no test form is under or over-represented in certain content areas).

SUMMARY

In 2008, the USMLE program administered approximately 140,000 Step or Step component examinations in the United States and around the world. One need only ponder this number to gain some appreciation for the labor and resources necessary to develop and maintain high quality examination content for the USMLE. The physicians and staff members associated with the program consider the USMLE the “gold standard” for medical licensing examinations. Maintaining this high standard remains a priority so that state medical boards may continue to rely upon the USMLE a quality independent assessment tool supplementing their judgment in the decision to grant an initial medical license.

*Non-overlapping refers to test forms whose scored content is not duplicated on another test form.

†Item discrimination reflects the differing performance on a question between individuals whose overall score is among the top half of examinees as opposed to examinees whose overall score is in the lower half. For a test question to be a good discriminator, most of the upper group should get the question right and most in the lower group should miss it.

ACKNOWLEDGEMENTS

Portions of this article previously appeared in the fall 2001 and spring 2006 issues of the Examiner, a publication of the National Board of Medical Examiners. The authors wish to thank the NBME for their permission to reprint material from the Examiner.

AUTHOR AFFILIATIONS

Ann Rhodes, B.S; David B. Swanson, Ph.D., Vice President, Professional Services, National Board of Medical Examiners; Katherine Z. Holtzman, Assistant Vice President, Test Development Services, National Board of Medical Examiners; David A. Johnson, M.A., Vice President, Assessment Services, Federation of State Medical Boards.

APPENDIX 1: STANDARD SETTING

The USMLE program provides a recommended pass or fail outcome on all Step examinations with numeric scores reported for Step 1, Step 2 CK and Step 3 in the form of a two- and three-digit scaled score. The recommended performance standards for the USMLE are based on a specified level of proficiency identified through a standard setting process. As a result, no predetermined percentage of examinees will pass or fail the examination.

Approximately every three years, each Step committee revisits its standard, i.e., minimum pass score. In discussing the appropriateness of the current standard, Step committees consider information drawn from multiple sources:

• recommendations from independent groups of physicians who have participated in content-based standard-setting activities;
• survey results from various groups such as state medical boards, medical school faculty, and examinees;
• trends in aggregate examinee performance data; and
• data on score precision and its effect on the pass/fail decision.

The content-based standard setting activities offer an especially important piece of data. This process involves in-
dependent panels of content experts. Participants on these panels are drawn from the medical licensure and undergraduate and graduate medical education communities and typically have had no prior experience writing test content for the USMLE program.

The panels begin by reviewing previously used test questions and rendering a judgment about the likelihood of a minimally proficient candidate correctly answering these questions. The panels then receive data detailing actual examinee performance on these questions. This provides the opportunity for panelists to reassess their concept of a minimally proficient examinee and revise their estimates for projected performance. Having completed this first phase of the process (i.e., judgment, feedback), the panels are then asked to review a new set of questions and make performance projections for minimally proficient examinees. Using the data derived from this second round of assessment by the panelists, staff then prepares a tentative minimum passing score based upon these experts’ judgment for subsequent review and consideration by the respective Step committee.

In addition to the results from these standard setting panels, the Step committee also reviews the results of surveys previously sent to representatives from licensing boards, the medical education community and examinees. The surveys questioned respondents on acceptable and unacceptable failure rates. By adding in data showing trends in examinee performance for that Step as well as psychometric details on score precision, the Step committee is able to address the fundamental question before them: “Do these data suggest a need to change the current minimum pass score?” If the answer to this question is “no”, the Step Committee decides to maintain the current standard; if “yes,” the committee then decides how much to change the current minimum passing score.

APPENDIX 2: PARTICIPATING WITH USMLE
Maintaining the high quality of its “national faculty” is a priority for the USMLE program. Staff associated with the program maintains a candidate database of prospective potential appointees to USMLE committees. Preserving a strong presence of physicians with state medical board experience is considered critical. Since 2007, the FSMB and NBME have hosted an annual item-writing workshop for state board members and/or participating in the USMLE program should submit their curriculum vitae to David Johnson, FSMB Vice President for Assessment Services, at P.O. Box 619850, Dallas, Texas 75261-9850 or via email: djohnson@fsmb.org.

REFERENCES