Improving the National Board of Medical Examiners Internal Medicine Subject Exam for Use in Clerkship Evaluation

D. Michael Elnicki, MD, Dianne A. Lescisin, MHPE, Susan Case, PhD

OBJECTIVE: To provide a consensus opinion on modifying the National Board of Medical Examiners (NBME) Medicine Subject Exam (Shelf) to: 1) reflect the internal medicine clerkship curriculum, developed by the Society of General Internal Medicine (SGIM) and the Clerkship Directors in Internal Medicine (CDIM); 2) emphasize knowledge important for a clerkship student; and 3) obtain feedback about students' performances on the Shelf.

DESIGN: Two-round Delphi technique.

PARTICIPANTS: The CDIM Research and Evaluation Committee and CDIM members on NBME Step 2 Committees.

MEASUREMENTS: Using 1–5 Likert scales (5 = highest ratings), the group rated test question content for relevance to the SGIM–CDIM Curriculum Guide and importance for clerkship students' knowledge. The Shelf content is organized into 4 physician tasks and into 11 sections that are generally organ system based. Each iteration of the Shelf has 100 questions. Participants indicated a desired distribution of questions by physician task and section, topics critical for inclusion on each exam, and new topics to include. They specified the types of feedback clerkship directors desired on students' performances. Following the first round, participants viewed pooled results prior to submitting their second-round responses.

RESULTS: Of 15 individuals contacted, 12 (80%) participated in each round. The desired distribution by physician task was: diagnosis (43), treatment (23), mechanism of disease (20), and health maintenance (15). The sections with the most questions requested were the cardiovascular (17), respiratory (15), and gastroenterology (12) sections. The fewest were requested in aging/ethics (4) and neurology, dermatology, and immunology (5 each). Examples of low-rated content were Wilson's Disease, chancroid and tracheal rupture (all <2.0). Health maintenance in type 2 diabetes, hypertension, and cardiovascular disease all received 5.0 ratings. Participants desired feedback by: section (4.6) and physician task (3.9), on performances of the entire class (4.0), and for individual students (3.8).

CONCLUSION: Clerkship directors identified test content that was relevant to the curricular content and important for clerkship students to know, and they indicated a desired question distribution. They would most like feedback on their students' performance by organ system–based sections for the complete academic year. This collaborative effort could serve as a model for aligning national exams with course goals.

KEY WORDS: clinical clerkship; clinical competence; student evaluation; undergraduate education.

J GEN INTERN MED 2002;17:435–440.

Evaluating medical students' knowledge is a critically important part of clinical clerkships. Because of its high reliability and national norms, many clerkship directors (CDs) use the National Board of Medical Examiners Subject Exam (Shelf) as a clerkship exam. In 1988, a national survey of internal medicine CDs found that 57% of those responding ($n = 101$) used the Shelf for a mean of 27% of students' grades.¹ From 1992 to 1998, the use of the Shelf steadily increased from 72% to 85% in internal medicine clerkships nationally.² The most recent survey of internal medicine CDs found that 83% use the Shelf for a mean of 24% of students' grades.³

The same survey demonstrated that 94% of CDs who used it were generally satisfied with the Shelf.³ However, some CDs express dissatisfaction with aspects of the exam, stating that it emphasizes unusual diseases and includes content areas not covered in their internal medicine curricula. Most CDs are quite familiar with the Shelf content, since the National Board of Medical Examiners (NBME) regularly asks CDs to review its content and to provide the NBME with feedback. The general content of the Shelf has also been discussed at national educational meetings, such as the Clerkship Directors in Internal Medicine (CDIM) annual meeting.

In 1997, the Society of General Internal Medicine (SGIM) and the CDIM jointly developed and disseminated a model curriculum for the internal medicine clerkship.⁴ The SGIM–CDIM curriculum focuses on generalist competencies and was largely derived by consensus among medical educators, including many CDs.⁵ It is increasingly being used as the curricular core of internal medicine clerkships,⁶ but the extent to which the Shelf matches this curriculum has not been formally assessed.

The NBME has expressed a willingness to work with CDs on improving the utility of the Shelf for CDs, which requires a consensus on what changes are required. The Delphi technique is a systematic method of gathering opinions on a topic through the use of sequential surveys, which are augmented by summary data and feedback.⁷ Initial opinions are distributed in an anonymous format to participants, who may modify their initial responses after reviewing the pooled responses of the others. The process requires time, motivation, and content expertise, but it has been successfully used to develop curricula as diverse as ambulatory clerkships⁸ and thoracic surgery residencies.⁹

Received from the University of Pittsburgh (DME, DAL), Pittsburgh, Pa; and the National Conference of Bar Examiners (SC), Madison, Wis.

Address correspondence and requests for reprints to Dr. Elnicki: Section of General Internal Medicine, UPMC Shadyside, University of Pittsburgh, 5230 Center Ave., Pittsburgh, PA 15232 (e-mail: elnickim@msx.upmc.edu).

We sought to provide a consensus opinion among a group of CDs, familiar with both the SGIM-CDIM Curriculum and the Shelf, on improving the exam for clerkship use and on providing feedback to the CDs about students' performances on the Shelf. The project was undertaken as a collaboration between the CDIM and the NBME.

METHODS

The participants in the Delphi process were the members of the CDIM's Research and Evaluation Committee and CDIM members who were currently serving on the NBME's Step 2 Committees, which are responsible for writing the internal medicine questions for the United States Medical Licensing Examination (USMLE) Step 2 Examination. They were informed of the purpose and the format of the exercise prior to agreeing to participate. Since little new content needed to be generated, it was felt that a 2-step Delphi would be sufficient to reach a consensus.

Each participant was sent a list of 357 keyword phrases, each of which described a test question from a recent Shelf exam. There are 100 questions on each Shelf exam. The keyword phrases are written to provide a good indication of the question content while maintaining the security of the examination. The keyword phrases are organized in two ways. The first is into 11 sections, which are based generally on organ systems and are listed in Table 1. The second organizational level is by 4 physician tasks (diagnosis, management, health maintenance, and mechanism of disease). Thus, an example of a keyword

phrase would be diagnosis of aortic stenosis, which would be in the cardiovascular section.

In the first round of the Delphi process, the group was asked to complete four tasks. The first task was to recommend the number of questions for each organ system-based section. The next task was to recommend the number of questions per exam based on physician task. The third task was to rate the Shelf content for "relevance to the SGIM-CDIM curriculum" and for "importance for clerkship students' knowledge." The latter designation was knowledge gained during the clerkship that was felt to be important for their development as physicians. The distinction between the two ratings was discussed prior to beginning the Delphi process. The ratings were done using 1–5 Likert scales with 1 = least and 5 = most. The Shelf content was evaluated at the level of physician task and section (e.g., diagnosis, immunologic disorders), but was not evaluated at the individual keyword phrase level on the first round. The final tasks were to generate lists of "critical diagnoses or problems" that should appear on each form of the Shelf exam and content areas that should be added to the Shelf exam.

The results of the first round were distributed to the group prior to their completing the second round, but the responses of individuals were kept anonymous. At this point, the members of the group were asked for feedback on the process and for other issues they wanted to address regarding the Shelf exam. Because many participants indicated that the sections did not allow sufficient detail to separate appropriate and inappropriate content, they asked to judge each keyword phrase individually in the second round. The first round also raised the issue of feedback on students' performances, so participants were asked to assess types of feedback that they desired on their students' performances in the second round.

The second round of the Delphi process again involved completing several tasks. The first was to rate each of 357 keyword phrases, using the "relevant" and "important" criteria as in round 1. The participants again were asked the preferred question distributions based on organ system sections and physician tasks. The third task was to rate the new topics that were suggested on the first round, using a 5-point scale with 1 = "don't add" to 5 = "must add." The fourth task was rating types of feedback on their students' performances as 1 = "don't want" to 5 = "must have." Participants were asked if they wanted feedback by individual student, complete class, section, or physician task. Finally, each of the critical topics was ranked by 1 = "strongly disagree" to 5 = "strongly agree."

The data analysis was performed using SPSS software (Version 10.0; SPSS Inc., Chicago, Ill). Means and standard deviations (SD) were calculated for each item. Correlations between relevance and importance ratings were calculated using Spearman's Rho. Comparisons of mean values, such as the numbers of questions by organ-based section or physician task were made using *t* tests, analysis of variance, or Mann-Whitney *U* test where appropriate.

Table 1. Number of Questions per 100-Question Shelf Exam that Clerkship Directors Desired by Organ System-Based Section*

Section	Mean	SD	95% CI
Cardiovascular disorders	17	2.5	15 to 18 [†]
Respiratory system diseases	15	2.0	14 to 17 [†]
Nutritional/digestive disorders	12	2.0	11 to 13 [†]
Renal, urinary, male reproductive systems disorders	11	1.4	10 to 12 [†]
Endocrine/metabolic disorders	11	1.6	10 to 12 [†]
Diseases of the blood/blood-forming organs	9	1.9	7 to 10
Musculoskeletal system/connective tissue disorders	8	1.7	7 to 9
Immunologic disorders	5	2.2	3 to 6 [‡]
Diseases of the nervous system and special senses	5	2.5	4 to 7 [‡]
Disorders of the skin and subcutaneous tissues	5	2.6	3 to 7 [‡]
Normal aging/ethics	4	1.4	3 to 5 [‡]

* The sections represent the basic organizational structure of the Shelf.

[†] *P* < .01 more than the expected *n* = 9 questions per exam;

[‡] *P* < .01 less than the expected *n* = 9 questions per exam.

CI, confidence interval.

RESULTS

Of the 15 eligible CDIM members contacted, 12 (80%) participated for each round of the Delphi process. Eleven participated in both rounds, 1 in only the first round and 1 in only the second round. Of the 13 responders, 5 (38%) were members of CDIM's Research and Evaluation Committee, 7 (54%) were on Step 2 Committees and 1 (8%) was on both. Results pertaining to the number of questions are expressed as whole numbers. Some totals do not equal 100 due to rounding.

In the first round, the participating CDs rated the cardiovascular and respiratory sections as the most relevant to the curriculum and important for clerkship students' knowledge (all means >4.0). They gave their lowest ratings (relevance and importance) to the dermatology (means 2.1 and 2.5) and neurology (means 2.1 and 2.6) sections. They asked for the following distribution of physician task (mean, SD): diagnosis (42, 6), treatment (24, 7), mechanism of disease (19, 5) and health maintenance (15, 5). The CDs asked that content on geriatric problems and women's health be added. They generated a

list of 42 critical topics that were rated in the second round.

The initial part of the second round dealt with the preferred distribution of questions by section. There are 11 sections in the Shelf, so the expected distribution of questions per examination would be 9 per section. The distribution of questions desired by the CDs is shown in Table 1. They requested disproportionately more questions in the cardiovascular, respiratory, gastrointestinal, endocrine, and renal sections. For the aging/ethics, immunology, neurology and dermatology sections, they desired fewer questions than expected on a random basis.

The desired question distribution by physician task in the second round by mean (SD) was: diagnosis 43 (7.8), therapy 23 (6.3), mechanism of disease 20 (4.2) and health maintenance 15 (6.0). None of these values were significantly different from the means generated in the first round.

The keyword phrases rated in the top 10% ($N = 31$) were rated ≥ 4.55 and are shown in Table 2. For each, the mean ratings of relevance and importance and SDs are shown. The correlations between relevance and importance were generally high with a mean $r = .70$. The critical topic list is

Table 2. Means, Standard Deviations (SD) and Corresponding Correlations (r) for the Keyword Phrases Rated Highest (Top 10%) by the Delphi Group

Keyword Phrase	Relevance Mean Rating (SD)	Importance Mean Rating (SD)	r Value
Cardiovascular disease, health maintenance	5.0 (0)	5.0 (0)	1.0
Coronary artery disease, health maintenance	5.0 (0)	5.0 (0)	1.0
Hypertension, diagnosis	5.0 (0)	5.0 (0)	1.0
Hypertension, health management	5.0 (0)	5.0 (0)	1.0
Diabetes mellitus, type 2, health maintenance	5.0 (0)	5.0 (0)	1.0
Diabetes mellitus, type 1, health maintenance	5.0 (0)	4.9 (.29)	1.0
Hypercholesterolemia, health maintenance	4.9 (.29)	4.9 (.29)	1.0
Heart failure, congestive, diagnosis	4.8 (.8)	4.9 (.29)	.52
Hypertension, management	4.8 (.62)	4.8 (.39)	1.0
Infarction, myocardial, mechanism	4.8 (.45)	4.8 (.39)	.78
Ischemia, myocardial, diagnosis	4.7 (.65)	5.0 (0)	1.0
Atherosclerosis, diagnosis	4.7 (1.16)	4.8 (.52)	.46
Angina pectoris, management	4.7 (.65)	4.8 (.45)	.59
HIV infection, diagnosis	4.7 (.49)	4.8 (.45)	.49
Diabetes mellitus, type 2, management	4.7 (.47)	4.8 (.45)	1.0
Asthma, management	4.7 (.49)	4.8 (.39)	.16
Cancer, colon, diagnosis	4.7 (.47)	4.8 (.45)	.77
Heart failure, congestive, management	4.7 (.65)	4.7 (.49)	.81
Nephropathy, diabetic, health maintenance	4.7 (.47)	4.6 (.51)	.81
Renal failure, diagnosis	4.6 (.69)	4.8 (.45)	.09
Pneumonia, management	4.6 (.51)	4.8 (.39)	.53
Pancreatitis, diagnosis	4.6 (.82)	4.6 (.51)	.80
Edema, pulmonary, mechanism	4.6 (.93)	4.5 (.67)	.42
Fibrillation, atrial, diagnosis	4.5 (.62)	4.9 (.29)	.60
Hyperosmolar coma, diagnosis	4.5 (.82)	4.8 (.45)	.61
Infarction, myocardial, inferior wall, diagnosis	4.5 (.82)	4.6 (.69)	.58
Anemia, iron deficiency, management	4.5 (.52)	4.6 (.51)	.85
Cancer, lung, diagnosis	4.5 (.9)	4.6 (.67)	.65
Fibrillation, ventricular, diagnosis	4.4 (.79)	4.8 (.39)	.55
Embolism, pulmonary, diagnosis	4.3 (1.23)	4.8 (.39)	.30
Embolism, pulmonary, health maintenance	4.3 (.98)	4.5 (.67)	.43

Note: Three hundred fifty-seven keyword phrases make up the Shelf content. Ratings were based on a 5-point scale where 1 = least relevant/important and 5 = most relevant/important. Relevance referred to the relevance of a keyword phrase to the SGIM-CDIM Curriculum. Importance referred to the importance of the subject represented by a keyword phrase to clerkship students' knowledge.

shown in Table 3, and there is clear similarity to the highly rated keyword phrases in Table 2. All but 8 of the critical topics received ratings ≥ 4.0 , and 84% (26/31) of the highly rated keyword phrases also appear on the critical topic list.

Those keyword phrases rated the lowest each had ratings ≤ 2.10 and are shown in Table 4. As with the highest rated, the mean ratings of relevance and importance, SDs, and correlations between relevance and importance are shown. The correlations between relevance and importance for each keyword phrase were again high with a mean $r = .66$. The low-rated items do not cluster

into any particular section, and none appear on the critical topics list.

Four types of feedback on student performance were rated for desirability. The highest rating was for receiving feedback by organ system-based section at a mean 4.6 (SD 0.52). Feedback by question category was rated 3.9 (SD 1.3). The desirability of receiving feedback at the individual student or complete class level was similar at 3.8 (SD 0.98) and 4.0 (SD 1.0), respectively. The new content areas suggested in round 1 were only moderately supported, in that geriatric problems was rated 3.6 (SD 0.51) and women's health 3.5 (SD 1.2).

Table 3. Means and Standard Deviations (SD) of Ratings Given to Critical Topics by the Delphi Group

Topic	Mean	SD
Coronary heart disease	4.91	0.30
Anemia	4.82	0.40
Congestive heart failure	4.82	0.40
Hypertension	4.82	0.40
Renal failure	4.82	0.40
Diabetes	4.82	0.40
Chronic obstructive lung disease	4.73	0.47
Asthma	4.73	0.47
Gastrointestinal bleed	4.70	0.67
Electrolyte abnormalities	4.64	0.50
Lipids	4.64	0.67
Peptic ulcer disease	4.60	0.70
Atrial fibrillation	4.55	0.69
Respiratory failure	4.55	0.69
Pneumonia	4.55	1.0
Breast cancer	4.50	0.85
Hypotension	4.45	0.52
Hepatitis	4.45	0.82
Alcoholic liver disease	4.45	0.69
Colon cancer	4.45	0.82
Acid-base disorders	4.45	0.69
Dyspnea	4.36	0.81
Human immunodeficiency virus	4.36	0.81
Thyroid disease	4.27	1.0
Upper respiratory infection	4.27	0.65
Electrocardiogram interpretation	4.18	0.75
Lung cancer	4.09	1.0
Pancreatitis	4.09	1.2
Gall bladder disease	4.09	0.94
Urinary tract infection	4.00	1.3
Syncope	4.00	0.89
Ethics	4.00	0.89
Cancer prevention	4.00	0.89
Cough	4.00	1.0
Arthritis	3.91	0.94
Headache	3.91	0.83
Delirium	3.82	0.75
Tuberculosis	3.82	0.75
Leukemia/lymphoma	3.64	0.92
Inflammatory bowel disease	3.64	0.81
Domestic violence	3.64	0.67
Portal hypertension	3.64	1.0

Note: The list of critical topics was generated by the Delphi group in the first round and rated in the second round. Ratings were based on a 5-point scale where 1 = strongly disagree and 5 = strongly agree that the topic is critical (i.e., should be on each form of the Shelf).

DISCUSSION

This study demonstrates that a group of CDs can articulate the content they believe is important to evaluate the knowledge bases of internal medicine clerkship students. Their selections of important topics closely mirror the clinical competencies from the national survey used to construct the CDIM-SGIM Curriculum. For example, the curriculum's highest rated ambulatory competency was hypertension, and the Delphi group (Appendix A) gave the diagnosis, health maintenance, and management of hypertension ratings of 5.0, 5.0 and 4.8, respectively.⁴ The high correlation between the relevance and importance questions for each keyword phrase reflects consistency in the responses.

There were few requests for new content to be inserted into the Shelf. The two areas that were suggested, geriatric disorders and women's health, received only measured support in the second round. This ambivalence is reflected in the Delphi group's asking for fewer questions on the normal aging section while asking for geriatric disease questions. However, the group may have distinguished between normal physiology and pathologic processes.

The low-rated items from the Shelf content do not appear in the SGIM-CDIM Curriculum. They were given low ratings for both relevance to the curriculum and importance for the students to know. Many of the lowest rated (e.g., Wilson's Disease, trachoma, chloramphenicol-induced anemia) are uncommon conditions, not likely to be encountered by medical students. Whether they should remain on the Shelf content list is a reasonable issue. However, the broad scope of the Shelf encourages students to learn about medicine beyond the limited and sometimes idiosyncratic events they experience in their clerkships. Some diseases may be rare but still worth students' attention because they demonstrate generalizable pathophysiologic processes or public health issues.

The use of NBME Shelf exams is a national trend in many clinical disciplines. Shelf scores have become a measure of both students' mastery of knowledge and courses' success at transferring that knowledge. Reflecting its widespread acceptance, the Shelf is also used to assess the effects of innovations in clerkships. It has been used to measure the effects of such diverse aspects as the adding of

Table 4. Means, Standard Deviation (SD) and Corresponding Correlations (*r*) for the Keyword Phrases Rated Lowest (Bottom 10%) by the Delphi Group

Keyword Phrase	Relevance Mean Rating (SD)	Importance Mean Rating (SD)	<i>r</i> Value
Cryptosporidiosis, mechanism	2.3 (1.06)	1.9 (.67)	.58
Immunodeficiency, IgA, mechanism	2.0 (.95)	2.2 (.94)	.37
Lymphoma, T-cell, cutaneous, diagnosis	2.0 (1.21)	2.1 (1.38)	.59
Paget's disease, mechanism	2.0 (.89)	2.1 (.90)	.61
Bite, spider, brown recluse, diagnosis	1.9 (1.31)	2.3 (.98)	.85
Infarction, splenic, diagnosis	1.9 (.83)	2.3 (.97)	.89
Angioedema, inherited, mechanism	1.9 (.90)	2.1 (.90)	.57
Molluscum contagiosum, mechanism	1.9 (.83)	1.9 (.79)	.93
Deficiency, vitamin D, mechanism	1.9 (1.30)	1.8 (.83)	.87
Impetigo, mechanism	1.8 (.75)	2.3 (.78)	.39
Acoustic neuroma, diagnosis	1.8 (.87)	2.3 (.78)	.77
Malaria, health maintenance	1.8 (.87)	2.1 (1.0)	.18
Coccidioidomycosis, diagnosis	1.8 (1.03)	2.1 (.90)	.84
Pityriasis rosea, diagnosis	1.8 (1.08)	2.0 (.95)	.88
Syringomyelia, diagnosis	1.8 (.98)	1.9 (1.08)	.77
Rocky Mountain spotted fever, management	1.7 (1.01)	2.3 (.97)	.75
Botulism, diagnosis	1.7 (1.01)	2.0 (1.04)	.63
Multiple endocrine neoplasia type 1, mechanism	1.7 (.79)	2.0 (.81)	.64
Nerve injury, axillary, mechanism	1.7 (1.01)	1.9 (1.00)	.85
Anemia, chloramphenicol-induced, mechanism	1.7 (.98)	1.7 (.65)	.66
Erythema multiforme, mechanism	1.6 (.67)	2.2 (1.03)	.45
Dermatitis, phototoxic, diagnosis	1.6 (.81)	2.0 (.85)	.86
Mucormycosis, diagnosis	1.6 (.90)	2.0 (1.13)	.71
Poisoning, liquid drain cleaner, diagnosis	1.6 (1.04)	1.8 (.94)	.73
Chancroid, mechanism	1.6 (.69)	1.6 (.79)	.97
Rhinitis, aspirin-induced, mechanism	1.5 (.67)	2.0 (.74)	.53
Wilson's disease, mechanism	1.5 (.67)	1.9 (.90)	.51
Rupture, tracheal, diagnosis	1.5 (.80)	1.8 (.97)	.44
Pyoderma gangrenosum, diagnosis	1.3 (.47)	1.8 (.58)	.23
Wilson's disease management	1.3 (.47)	1.6 (.67)	.72
Trachoma, diagnosis	1.3 (.62)	1.6 (.67)	.59

Note: Three hundred fifty-seven keyword phrases make up the Shelf content. Ratings were based on a 5-point scale where 1 = least relevant/important and 5 = most relevant/important. Relevance referred to the relevance of a keyword phrase to the SGIM-CDIM Curriculum. Importance referred to the importance of the subject represented by a keyword phrase to clerkship students' knowledge.

a didactic curriculum,¹⁰ an ambulatory block rotation,¹¹ and a standardized patient examination to a clerkship.¹² The Shelf has also been used to identify students at risk for failing the USMLE Step 2 Exam¹³ and to compare the performances of students with ambulatory versus inpatient experience.¹⁴ Its use may reflect either a high regard for the Shelf or a lack of time for faculty to write multiple-choice exams.¹⁵

The CDs would like to have feedback on their students' performance at the class level and by organ system. Their desire for this type of feedback probably reflects the value that CDs place on nationally normed scores. This type of information helps CDs to adjust their curriculum content and delivery methods in areas of poor performance. It also enables them to congratulate those involved in areas of high performance. The CDs' desire for feedback needs to be tempered by the realization of reliability issues. As the number of items in a content area decreases, the uncertainty of the subscores increases. There will be a limited number of questions on an organ system, which may lead to inaccurate conclusions about students' performances.

There are several strengths in the design of the study. The Delphi instrument was constructed with input from both the NBME and the CDIM. The Delphi group was highly motivated, as demonstrated by the request for a more-detailed second round. The group members had intimate knowledge of both the Shelf and the CDIM-SGIM Curriculum designs. The group members constituted a diverse, national representation of CDs with an interest in evaluation techniques.

There are limitations to this study, which need to be acknowledged. As in any cross-sectional survey, the opinions expressed are those for single points in time. Areas deemed important in students' curricula will inevitably change with time. The opinions of the members of the Delphi group may not be representative of CDs nationally. Many of the participating CDs have been involved in the developing and disseminating of the SGIM-CDIM Curriculum, and their investment may have inflated the correlation between its content and knowledge they think is important for clerkship students. Lastly, the Delphi process favors consensus and may minimize the impact of opinions held by a minority.

Multiple-choice examinations remain the standard method of assessing students' knowledge bases. They are efficient, relatively inexpensive, and when carefully prepared, as are the Shelf exams, highly reliable. However, they do not give a complete picture of a student's competency. Multiple-choice exams do not measure clinical skills or interactive abilities well.¹⁶ Some of these limitations are intrinsic to the exam style, but a step could be taken toward improving the content validity of the Shelf by aligning it with core curriculum content. We hope that this study will lead to closer collaboration between CDs and the NBME. This process could serve as a model for aligning exams and curriculum at a national level.

The authors wish to thank Sylvia Ford for preparing the manuscript.

REFERENCES

1. Magarian GJ, Mazur DJ. Evaluation of students in medicine clerkships. *Acad Med.* 1990;65:341-5.
2. Kassebaum DG, Eaglen RH. Shortcomings in the evaluation of students' clinical skills and behaviors in medical school. *Acad Med.* 1999;74:842-9.
3. Hemmer PA, Elnicki DM, Szauder K, et al. How do medicine clerkship directors use and view the National Board of Medical Examiners Subject Examination in Medicine. Presented at the 39th Annual Research in Medical Education Conference, Chicago, Ill, October, 2000.
4. Bass EB, Morrison G, Wills S, Mumford LM, Goroll AH. National survey of clerkship directors in internal medicine on the competencies that should be addressed in the medicine core clerkship. *Am J Med.* 1997;102:564-71.
5. Goroll AH, Morrison G, Bass EB, et al. Reforming the Core Clerkship in Internal Medicine: The SGIM/CDIM Project. *Ann Intern Med.* 2001;134:30-7.
6. Jablonover RS, Blackman DJ, Bass EB, Morrison G, Goroll AH. Evaluation of a national curriculum reform effort for the medicine core clerkship. *J Gen Intern Med.* 2000;15:484-91.
7. Murry JW, Hammons JO. Delphi: a versatile methodology for conducting qualitative research. *Review of Higher Education.* 1995;18:423-36.
8. Stritter FT, Tresolini CP, Reeb KG. The Delphi technique in curriculum development. *Teach Learn Med.* 1994;6:136-41.
9. Wilcox BR, Stritter FT, Anderson RP, et al. Systematic survey of opinion regarding the thoracic surgery residency. *Ann Thorac Surg.* 1993;55:1296-302.
10. Magarian GJ. Influence of a medicine clerkship conference series on students' acquisition of knowledge. *Acad Med.* 1993;68:923-6.
11. Pangaro L, Gibson K, Russell W, Lucas C, Marple R. A prospective, randomized trial of a six-week ambulatory medicine rotation. *Acad Med.* 1995;70:537-41.
12. Elnicki DM, Shockcor WT, Morris DK, Halbritter KA. Creating an objective structured clinical examination for the internal medicine clerkship: pitfalls and benefits. *Am J Med Sci.* 1993;306:94-7.
13. Ripkey DR, Case SM, Swanson DB. Identifying students at risk for poor performance on USMLE Step 2. *Acad Med.* 1999;74(10 suppl):45-8.
14. Fincher RE, Case SM, Ripkey DR, Swanson DB. Comparison of ambulatory knowledge of third-year students who learned in ambulatory settings with that of students who learned in inpatient settings. *Acad Med.* 1997;72(suppl):130-2.
15. Mavis BE, Cole BL, Hoppe RB. A survey of student assessment in U.S. medical schools: the balance of breadth versus fidelity. *Teach Learn Med.* 2000;13:74-9.
16. Elnicki DM, Ainsworth MA, Magarian GJ, Pangaro LN. Evaluating the internal medicine clerkship. *Am J Med.* 1994;97:I-VI.

APPENDIX A

Members of the Delphi Group

T. Andrew Albritton, MD, Medical College of Georgia
 Michael Battistone, MD, University of Utah
 Lynn Cleary, MD, SUNY Upstate School of Medicine
 Gary Ferenchick, MD, Michigan State University
 Charles H. Griffith, III, MD, University of Kentucky
 Cyril Grumm, MD, University of Michigan
 Paul A. Hemmer, MD, Uniformed Service University of the Health Sciences
 Walter Kernan, MD, Yale University
 Thomas Painter, MD, University of Pittsburgh
 Mark M. Udden, MD, Baylor College of Medicine
 Paul Wallach, MD, University of South Florida
 Allison Whelan, MD, Washington University School of Medicine
 Raymond Wong, MD, Loma Linda University