Development, Implementation, and Results of the ASN In-Training Examination for Fellows

Mitchell H. Rosner,* Jeffrey S. Berns,† Mark Parker,‡ Ashita Tolwani,§ James Bailey,‖ Susan DiGiovanni,¶ Eleanor Lederer,** Suzanne Norby,†† Troy J. Plumb,‡‡ Qi Qian,†† Jane Yeun,§§ Janine L. Hawley,‖‖ and Susan Owens¶¶ (The ASN In-Training Examination Committee)

**University of Virginia Health System, Charlottesville Virginia; †University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania; ‡Maine Medical Center and Tufts University School of Medicine, Portland, Maine; §University of Alabama at Birmingham School of Medicine, Birmingham, Alabama; ‖Emory University School of Medicine, Atlanta, Georgia; ¶Virginia Commonwealth School of Medicine, Richmond, Virginia; **Louisville Veterans Administration Hospital and University of Louisville School of Medicine, Louisville, Kentucky; ††Mayo College of Medicine, Rochester, Minnesota; ‡‡University of Nebraska Medical Center, Omaha, Nebraska; §§University of California–Davis School of Medicine, Davis, California; ‖‖National Board of Medical Examiners, Philadelphia, Pennsylvania; and ¶¶American Society of Nephrology, Washington, DC*

The American Society of Nephrology and the fellowship training program directors in conjunction with the National Board of Medical Examiners developed a comprehensive assessment of medical knowledge for nephrology fellows in-training. This in-training examination (ITE) consisted of 150 multiple-choice items covering 11 broad content areas in a blueprint similar to the American Board of Internal Medicine certifying examination for nephrology. Questions consisted of case vignettes to simulate real-life clinical experience. The first examination was given in April 2009 to 682 fellows and six training program directors. Examinees felt that the examination was well structured and relevant to their training experience Longitudinal performance on the examination will be helpful in allowing training programs to utilize results from content areas in identifying deficits in medical knowledge as well as assessing the results of any curriculum changes.

Clin J Am Soc Nephrol 5: 328–334, 2010. doi: 10.2215/CJN.06860909

The Accreditation Council for Graduate Medical Education (ACGME) Outcome Project requires that fellowship training programs document that trainees demonstrate competence in six core areas: medical knowledge, patient care, practice-based learning and improvement, systems-based practice, professionalism, and interpersonal and communication skills (1). Evaluation of medical knowledge is critically important: this competency forms the foundation for clinical competence and serves as the basic requisite for the development of other clinical skills (2). Recognition of the importance of competency in medical knowledge has led to widespread use of in-training examinations (ITEs) for subject trainees (either residents or fellows) to assess their medical knowledge through multiple-choice questions (3,4). As opposed to the evaluation of other ACGME competencies, assessment of medical knowledge easily lends itself to a written examination that covers the relevant content areas of the specialty (5). A written examination can provide valid and reliable results that can be utilized in an ongoing program of improvement for the trainees and training program and can be statistically analyzed for internal validity and various other purposes. Along these lines, trainees have used their examination results to identify areas of deficiency that allow more focused study, compare their performance with that of their peers in training, and track their improvement over their training experience (6). Program directors have utilized these results to identify deficiencies in their trainees, provide educational guidance, and develop teaching programs in areas that have been identified as weak (6). Longitudinal scores on in-training examinations allow fellowship training program directors to assess and validate changes in curriculum (6).

The Training Program Director's Executive Committee of the American Society of Nephrology (ASN) began development of an ITE for fellows in 2006 on the basis of an initiative promoted by the Association of Specialty Professors. Dr. Mark Rosenberg led the initial effort in this regard. This report describes the development and implementation process of the examination, as well as the results from the first ITE administered in April 2009.

Development of the Examination

The ITE was developed with several goals in mind: (*1*) to provide training program directors with information regarding

Published online ahead of print. Publication date available at www.cjasn.org.

Correspondence: Dr. Mitchell H. Rosner, University of Virginia Health System, Division of Nephrology, Box 800133, Charlottesville, VA 22908. Phone: 434-924-2187; Fax: 434-924-5848; E-mail: mhr9r@virginia.edu

ISSN: 1555-9041/502–0328

the potential strengths and weaknesses in their individual programs through benchmarking of their program compared with other programs on the basis of scores on the examination, (*2*) to provide feedback to fellows in-training regarding their relative strengths and weaknesses in specific content areas as well as to allow them to compare their results against national outcomes, (*3*) to facilitate identification and discussion of potential weaknesses in education in nephrology training programs nationally, (*4*) to allow fellows to track changes in their medical knowledge between their first and second years of renal fellowship, and (*5*) to allow assessment of medical knowledge competency in trainees.

ASN contracted with the National Board of Medical Examiners (NBME), a vendor with extensive experience in the development, implementation, and evaluation of medical specialty examinations. The NBME also had the capability of building and delivering an electronically based examination.

An ITE steering committee was convened comprised of members of the ASN training program director executive committee. After this, NBME conducted a half-day question-writing workshop at the training program director retreat, where they instructed interested test item writers in the design of high-quality test questions (*7*). From this workshop, a group of volunteers was selected to form the Test Material Development Committee. A total of 12 nephrologists formed this committee with a broad array of expertise covering the entire spectrum of nephrology content. Each committee member was asked to submit questions in particular content domains. These content domains were chosen to mirror the blueprint of the American Board of Internal Medicine certifying examination for nephrology (Table 1). The questions were written in a case-based vignette format to be compatible with real-life clinical experiences. Where applicable, high-resolution digital images of radiographs and kidney biopsies were included. To ensure high quality, stylistic consistency, and clarity, the NBME staff reviewed the submitted questions. Items were rewritten as needed. After this review, the ITE committee reviewed a total of 300 items to select a final examination of 150 items meeting the blueprint requirements. Each question was reviewed and assigned to a primary content area (Table 1) and to a particular task code (diagnosis, management, outcome/prognosis, or to a basic science classification). The committee was also asked to grade each question (A, B, or C) on the basis of quality and ability to test key concepts. Only grade A or B questions were utilized for the examination. After the final 150 questions were selected, a subset of the committee reviewed the final form of the examination, selected 10 questions for a practice test, and performed a final review of the electronic (Internet) version of the examination. The ten practice questions were posted on the ASN website for trainees and program directors to review.

The examination was administered at 123 test centers under proctored conditions using a web-based delivery system on April 22 and 23, 2009. Examinees were allotted up to 6 hours to take the examination. A total of 682 examinees (301 first-year fellows and 381 second-year or higher fellows) took the examination. On the basis of the most recent estimates, there are 812 fellows training in nephrology (291 women, 322 U.S. medical graduates, 431 international medical graduates, and 49 osteopathic school graduates); thus, not all fellows in training took the examination (*8*). There are several possible reasons for this: (*1*) participation was voluntary, (*2*) some programs decided that first-year fellows may not have had enough experience to take the examination, and (*3*) cost of the examination ($240 per fellow) may have limited participation. The committee also asked training program directors if they were interested in taking the examination for quality assessment and improvement purposes; six program directors took the examination. A detailed postexamination survey was also performed that assessed the quality and content of the examination and demographic features of the fellow cohort.

Results and Analysis of the ITE

For the purposes of statistical analysis, a reference group of 293 second-year (or higher) fellows was arbitrarily defined to serve as a benchmark for data comparison. Each item response was analyzed, and content experts identified for review items with unusual statistical behavior ("key validation"). The criteria for identifying these items included (*1*) items for which

Table 1. Examination content

Subject Area	Percent of Examination Questions Devoted to Content Area
Chronic kidney disease	13
Glomerular/vascular disease	12
Tubulointerstitial/cystic disease	6
Acute renal failure/intensive care unit nephrology	10
Transplantation	10
Hypertension	10
Electrolytes: sodium and water	9
Electrolytes: potassium and acid-base	9
Mineral metabolism	8
Pharmacology	9
Ethics	4

fewer than 30% of the examinees responded correctly, (*2*) items that had a negative item discrimination index (indicating that low-scoring examinees responded correctly more frequently than high-scoring examinees), (*3*) a wrong answer was chosen more often than the keyed-in correct answer, (*4*) the high-performing examinees chose an alternative answer more often than the keyed-in correct answer, or (*5*) an alternative answer was chosen more frequently. Thirteen items were identified for review and discussed by NBME and content experts. As a result, four items were deleted from scoring and one item was re-keyed. The remaining nine items were determined to be correct and were included in the final scoring. Thus, 146 of 150 items on the examination were scored.

After key validation decisions were implemented, raw and scaled scores were computed for each fellow. The raw score is the number of items answered correctly on the test. Scaled scores are calculated such that the base reference group (293 second-year or higher fellows) scores had a mean of 500 and a SD of 100. Note that the mean score for second-year fellows is lower than 500 (495), reflecting the fact that a subgroup of 293 of 381 total second-year fellows was used for standardization. It should be noted that fellowship programs differ as to when clinical *versus* research years are allocated and whether a research fellowship is offered. Thus, in the base reference group, we do not know how many of these fellows were in research programs or have completed their clinical training. Similarly, in the first-year fellow group, we do not know how many of these fellows are in research tracks and may have had limited clinical exposure. A histogram that depicts the distribution of the scores for all examinees is shown in Figure 1. In addition to scores on the total examination, content area scores were calculated and scaled so that the base reference group scores had a mean of 70 and a SD of 8. This base reference group score distribution serves as a benchmark for comparison with future examination performance. Table 2 shows the percent of items answered correctly in content areas for first- and second-year fellows as well as for the entire group. In most content areas, as well as for the total examination, the performance of first-year fellows was similar or slightly inferior to second-year (or higher) fellows. This may reflect the fact that the examination was given toward the end of the first fellowship year after much didactic content was given to fellows and that advancing fellowship year may not necessarily correlate with increased clinical experience or medical knowledge attainment.

Table 3 shows the number of items scored (and deleted after key validation) and the item difficulty (*P* value) for each content area for the reference group of 293 second-year fellows. Item difficulty (*P* value) is the proportion of examinees that gave a correct response to an item. Item difficulty is influenced by the intrinsic difficulty of the item and the proficiency of examinees, and this statistic can be utilized to compare the difficulty of the examination across content areas. For this examination, the mean difficulty was 0.69, with content areas ranging from 0.58 (kidney transplantation) to 0.80 (mineral metabolism).

Examination scores are only an estimate of an examinee's true proficiency level, with the standard error of measurement (SEM) providing additional statistical information about the degree of uncertainty in the scores. In general, the more items on which a score is based, and the higher the discrimination of those items, the smaller the SEM. An examinee's score is likely (2 times out of 3) to be within 1 SEM of his or her "true" proficiency level. The average SEM for this year's examination was approximately 3% (equivalent to 46 scaled units). This means that if an examinee's true proficiency were 500, the score that he or she actually receives on the examination will likely (2 times out of 3) fall within 454 and 546.

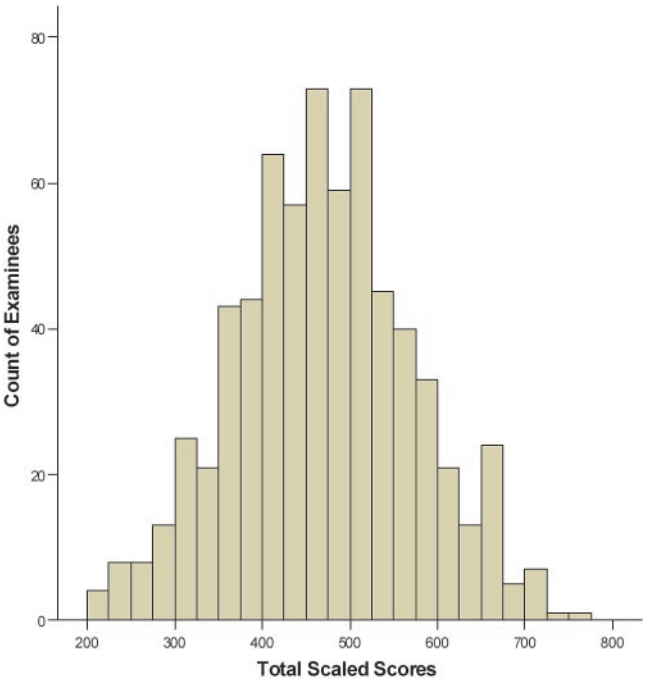

Figure 1. Histogram of total test scaled scores for the total group, where the total group consists of all candidates who took this administration of the ITE (mean = 469, SD = 103, $N = 682$).

Examination results for each fellow were sent to their respective program director as well as a summary of the performance of all fellows in all programs. A histogram was provided that showed the distribution of mean scores for all programs, for first- and second-year (or higher) fellows separately, as well as the mean score for their particular program. For each examinee, the results showed their total score and a score within each content area. Examinees could use these reports to determine where they stood relative to their peer group by year of fellowship. The report sent to training program directors did not include results of individual items or provide topics from individual questions. There were several reasons why this information was not provided: (*1*) the examination is not meant to have programs or trainees study specific topics but is designed to provide an overall assessment of medical knowledge competency; and (*2*) within each content area the SD can be high because the number of items are few, thus a content-specific score may be misleading.

A post-test survey was utilized to assess examinee feedback

Table 2. 2009 ASN ITE mean examinee performance (percent correct scores) by fellowship year[a]

Content Area	First-Year Fellows (n = 301)	Second-Year and Higher Fellows[b] (n = 381)	All Fellows (n = 682)
General aspects of chronic kidney disease	65 (12)	71 (11)	68 (12)
Glomerular/vascular disorders	57 (12)	61 (12)	60 (12)
Tubular/interstitial and cystic disorders	74 (16)	81 (16)	78 (17)
Acute renal failure/intensive care unit nephrology	61 (11)	63 (11)	62 (11)
Kidney transplantation	52 (14)	58 (15)	55 (15)
Hypertension	65 (13)	68 (12)	67 (12)
Electrolyte physiology: sodium/water	74 (14)	78 (13)	76 (13)
Electrolyte physiology: acid-base/potassium	57 (15)	61 (14)	59 (15)
Mineral metabolism	72 (16)	80 (13)	77 (15)
Clinical pharmacology	70 (13)	73 (13)	72 (13)
Ethics	75 (21)	76 (20)	76 (20)
Total test	64 (8)	69 (7)	67 (8)

[a]Values presented as mean (SD).
[b]This group also includes examinees in their third, fourth, and fifth year of fellowship.

Table 3. 2009 ASN ITE reference group[a] (n = 293) summary item statistics by content area

Content Area	Number of Items Scored (Deleted)	Item Difficulty (P value) Mean (SD)
General aspects of chronic kidney disease	18 (2)	0.71 (0.20)
Glomerular/vascular disorders	19 (0)	0.62 (0.26)
Tubular/interstitial and cystic disorders	8 (1)	0.80 (0.12)
Acute renal failure/intensive care unit nephrology	16 (0)	0.64 (0.28)
Kidney transplantation	14 (1)	0.58 (0.21)
Hypertension	15 (0)	0.68 (0.21)
Electrolyte physiology: sodium/water	13 (0)	0.78 (0.16)
Electrolyte physiology: acid-base/potassium	14 (0)	0.62 (0.20)
Mineral metabolism	11 (0)	0.80 (0.13)
Clinical pharmacology	14 (0)	0.73 (0.19)
Ethics	4 (0)	0.76 (0.20)
Total	146 (4)	0.69 (0.21)

[a]The reference group is composed of second-year fellows who took the examination under standard test conditions and who experienced few or no incidents on the testing day.

on the examination and obtain demographic information regarding nephrology fellows. Seventy-three percent of fellows felt that the examination was at the appropriate level of difficulty, 10% felt that the examination was too hard, and 0.4% felt that the examination was too easy (16% of fellows did not respond). Fellows were asked to judge whether the examination material was relevant or not: 41.7% responded that the examination was very relevant, 34.9% moderately relevant, 7.4% somewhat relevant, and 0.1% not at all relevant (15.8% of fellows did not respond). No subject content area was felt by fellows to be either overemphasized or underemphasized on the examination (defined as responses to this question for each content area <15%).

Subgroup Analysis of Scores

Fellows were asked to provide detailed information regarding the medical school they attended, size of their fellowship program, type and size of teaching hospital, career plans, and primary sources of medical information. Answers to these items were utilized to perform subgroup analysis with examination performance. It should be cautioned that such an analysis is based on not only self-reporting, but also on only one examination and sample sizes that, in some cases, are small. Statistical analysis of comparison between subgroups was performed with unpaired t-tests, with statistical significance defined as a P value <0.05.

As shown in Figure 2, scores based on the classification of the medical school that fellows graduated from reveals that with the exception of off-shore (Caribbean) U.S. medical schools, there was comparable performance across locations of medical school training. The number of off-shore U.S. medical graduates was significantly smaller than other groups (n = 17); however, the difference in scores between

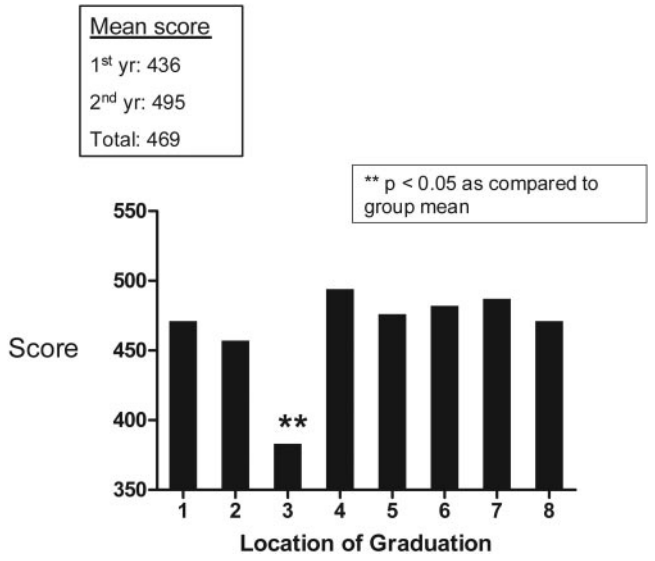

Figure 2. Examination scores based upon classification of medical school. 1 = U.S. medical school granting MD degree (n = 244), 2 = U.S. medical school granting DO degree (n = 40), 3 = U.S. medical school located off-shore (n = 17), 4 = international medical school located in Europe (n = 26), 5 = international medical school located in the Middle East (n = 45), 6 = international medical school located in Asia (n = 188), 7 = international medical school located in Latin America (n = 28), 8 = international medical school located in Africa (n = 19).

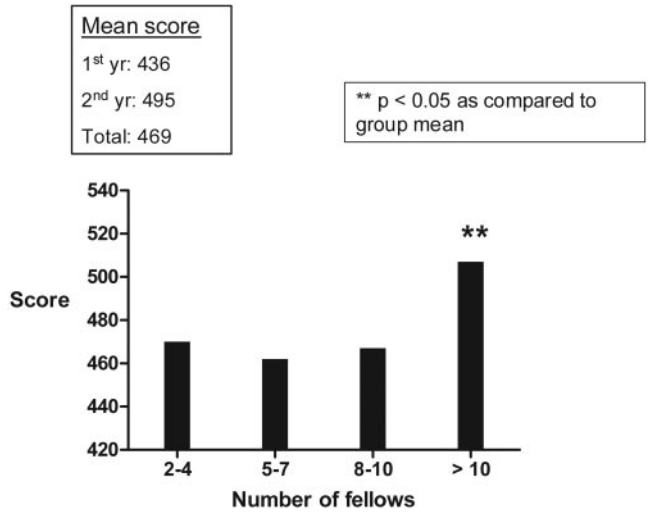

Figure 3. Examination scores stratified according to the number of fellows in the training program.

these examinees and the mean score was statistically significant ($P < 0.05$). This finding suggests that, for unclear reasons, off-shore medical graduates are less prepared for testing of medical knowledge. Given the small sample size and single time point, this finding will need to be followed in subsequent examinations.

When scores were stratified based on program size (*i.e.*, total number of fellows enrolled in the program), there was a statistically significant higher mean score (compared with the overall

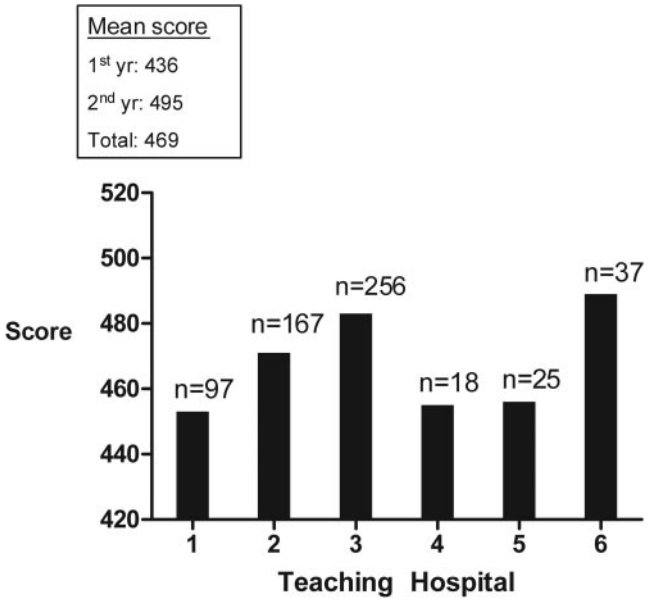

Figure 4. Examination scores stratified according to the size and type of major teaching hospital affiliated with the fellow training program. 1 = university medical center with 200 to 399 beds (n = 97), 2 = university medical center with 400 to 600 beds (n = 167), 3 = university medical center with >600 beds (n = 256), 4 = community medical center with 200 to 399 beds (n = 18), 5 = community medical center with 400 to 600 beds (n = 25), 6 = community medical center with >600 beds (n = 37).

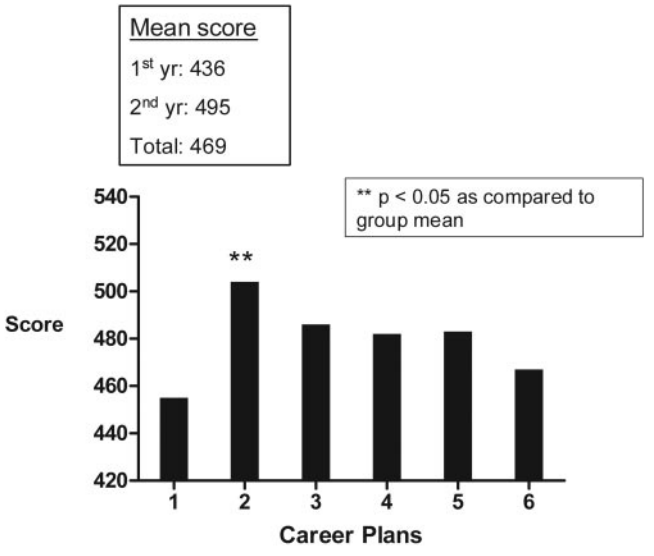

Figure 5. Examination scores stratified according to stated career plans of fellow. 1 = no response (n = 95), 2 = academic investigator (spending majority of time performing basic science research) (n = 31), 3 = clinical investigator (spending majority of time performing clinical research) (n = 43), 4 = clinician educator (spending time divided between teaching and clinical care) (n = 85), 5 = academic clinician (spending majority of time in clinical care at an academic medical center) (n = 79), 6 = private practice (n = 325).

Table 4. Primary sources of medical information for fellow trainees

Medical Information Source	Percent of Fellows Selecting this Source as One of Their Top Three Sources of Medical Information
UpToDate	76.6
Journal review articles	57.1
Textbooks	41.4
ASN Nephrology Self-Assessment Program	22.8
Original scientific articles	21.5
Internet searches	17.4
Nephrology-specific websites	5.4

group mean score) in programs with more than ten fellows (Figure 3). No difference in mean scores was noted between programs with sizes less than ten fellows. One possible reason for this association is that programs with more than ten fellows may be more likely to have a higher proportion of fellows beyond the first year of training. The group of fellows in their second or higher year of training achieved higher scores on the ITE than first-year fellows overall. Scores based on the major teaching hospital utilized in the fellowship program (Figure 4) revealed a trend for higher examination scores in larger medical centers (either university or community) that did not reach statistical significance. When analyzed on an examination-content-area basis, the trend for larger medical centers having higher scores was seen in all content areas and not specific to one area.

Fellow examinees were asked to state potential career plans and these data were used to stratify exam results (Figure 5). With the exception of fellows stating that they planned to pursue a career as an academic investigator (defined as most of their time devoted to basic science), who performed significantly better than the overall group mean ($P < 0.05$), median scores across career plan groups was not statistically different from the group mean. One possible explanation for this finding was that fellows who identified themselves as interested in a career in academic investigation were more likely to be second-year or higher fellows; however, subgroup analysis revealed that this was not the case. A limitation in this analysis is the large number of fellows (95) who did not respond with a career choice. This subgroup had the lowest mean examination score and if this group included fellows who were also interested in academics, then the mean score would be lowered toward other career plan subgroups.

Finally, fellows were asked to rank their top three sources of information for nephrology. The highest ranked information source was UpToDate (Boston, MA) (76.6%; Table 4). Fellow scores stratified by which source of information was ranked highest showed no differences across the various sources listed in Table 4 (data not shown).

Six training program directors also took the examination. Their performance placed in the top 5% of examinees. These program directors were surveyed regarding their examination experience and 100% felt that the examination was comprehensive and at the appropriate level of difficulty. All of these program directors stated that they would recommend that all trainees take the ITE.

In summary, the ASN developed an ITE for nephrology fellows over a 3-year period in response to a need for training program directors to provide an objective assessment of trainee medical knowledge. The examination also allows training program directors to benchmark their program against others and determine if deficiencies in their training program exist in any specific content areas. Fellows will be able to use their examination results over consecutive years to assess progress in their own medical knowledge and guide their preparation for board examination preparation. Most examinees felt that the examination was fair and covered appropriate topics. The electronic format was well received and functioned well. The statistical performance of the examination was excellent. As more cohorts of trainees take the examination, it will be possible to track performance of individual trainees and training programs over time and correlate outcomes with the results of the American Board of Internal Medicine certifying examination. This will allow a better assessment of any deficiencies in content areas and help formulate larger initiatives to educate fellows in these areas. The success of the development and implementation of the ITE was due to a collaborative effort between the ASN leadership, training program directors, and the ITE development committee.

Acknowledgments

The ITE Committee thanks all of the training program directors for their dedication to this project. The ITE Committee also thanks Tod Ibrahim, the ASN council, and the dedicated staff at the NBME for their support of this project.

Disclosures

None.

References

1. Accreditation Council for Graduate Medical Education: Outcome Project. Available online at http://www.acgme.org/outome/comp/refList.asp. Accessed September 10, 2009
2. Hawkins RE, Swanson DB. Using written examinations to assess medical knowledge and its application. In: *Practical Guide to the Evaluation of Clinical Competence*, edited by

Holmboe ES, Hawkins RE, Philadelphia, Elsevier Health Sciences, 2008: pp 42–59
3. American College of Physicians: Internal Medicine In-Training Exam. Available online at http://www.acponline.org/education_recertification/education/in_training/. Accessed September 28, 2009
4. Collichio FA, Kayoumi KM, Hande KR, Hawkins RE, Hawley JL, Adelstein DJ, D'Angelo JM, Stewart JA: Developing an in-training examination for fellows: The experience of the American Society of Clinical Oncology. *J Clin Oncol* 27: 1706–1711, 2009
5. Epstein RM: Assessment in medical education. *N Engl J Med* 356: 387–396, 2007
6. Garibaldi RA, Subhiyah R, Moore ME, Waxman H: The in-training examination in internal medicine: An analysis of resident performance over time. *Ann Intern Med* 137: 505–510, 2002
7. Case SM, Swanson DB: Constructing Written Test Questions for the Basic and Clinical Sciences, 3rd ed.: Philadelphia, National Board of Medical Examiners, 2002
8. Brotherton SE, Etzel SI: Graduate medical education, 2008–2009. *JAMA* 302: 1357–1372, 2009