Endobronchial Ultrasound Skills and Tasks Assessment Tool

Assessing the Validity Evidence for a Test of Endobronchial Ultrasound-guided Transbronchial Needle Aspiration Operator Skill

Mohsen Davoudi[1], Henri G. Colt[1], Kathryn E. Osann[1], Carla R. Lamb[2], and John J. Mullon[3]

[1]Pulmonary and Critical Care Division, Department of Medicine, University of California Irvine, Irvine, California; [2]Pulmonary and Critical Care Department, Lahey Clinic, Burlington, Massachusetts; and [3]Pulmonary and Critical Care Medicine, Mayo Clinic, Rochester, Minnesota

***Rationale*: Endobronchial ultrasound-guided transbronchial needle aspiration (EBUS-TBNA) is becoming standard of care for the sampling of mediastinal adenopathy. The need for a safe, effective, accurate procedure makes EBUS-TBNA ideal for mastery training and testing.**

***Objectives*: The Endobronchial Ultrasound Skills and Tasks Assessment Tool (EBUS-STAT) was created as an objective competency-oriented assessment tool of EBUS-TBNA skills and knowledge. This study demonstrates the reliability and validity evidence of this tool.**

***Methods*: The EBUS-STAT objectively scores the EBUS-TBNA operator's skills, including atraumatic airway introduction and navigation, ultrasound image acquisition and optimization, identification of mediastinal nodal and vascular structures, EBUS-TBNA sampling, and recognition of EBUS/computed tomography images of mediastinal structures. It can be administered at the bedside or using combination of low- and high-fidelity simulation platforms. Two independent testers administered the EBUS-STAT to 24 operators at three levels of EBUS-TBNA experience (8 beginners, 8 intermediates, and 8 experienced) at three institutions; operators were also asked to self-assess their skills. Scores were analyzed for intertester reliability, correlation with prior EBUS-TBNA experience, and association with self-assessments.**

***Measurements and Main Results*: Intertester reliability between testers was very high ($r = 0.9991$, $P < 0.00005$). Mean EBUS-STAT scores for beginner, intermediate, and experienced groups, respectively, were 31.1, 74.9, and 93.6 out of 100 ($F_{2,21} = 118.6$, $P < 0.0001$). Groups were nonoverlapping: *post hoc* tests showed each group differed significantly from the others ($P < 0.001$). Self-assessments corresponded closely to actual EBUS-STAT scores ($r^2 = 0.81$, $P < 0.001$).**

***Conclusions*: The EBUS-STAT can be used to reliably and objectively score and classify EBUS-TBNA operators from novice to expert. Its use to assess and document the acquisition of knowledge and skill is a step toward the goal of mastery training in EBUS-TBNA.**

Keywords: bronchoscopy; education; competency; endobronchial ultrasound-guided transbronchial needle aspiration; assessment

AT A GLANCE COMMENTARY

Scientific Knowledge on the Subject

Endobronchial ultrasound-guided transbronchial needle aspiration (EBUS-TBNA) is now the standard of care for mediastinal sampling. Yet, there is currently no assessment tool to test the competencies of EBUS-TBNA operators.

What This Study Adds to the Field

We demonstrate that the Endobronchial Ultrasound Skills and Tasks Assessment Tool can be used to reliably and objectively score and classify EBUS-TBNA operators from novice to expert.

Endobronchial ultrasound-guided transbronchial needle aspiration (EBUS-TBNA) is now standard of practice for the sampling of mediastinal and hilar adenopathy for suspected malignancy as well as for staging primary or metastatic cancer involving the mediastinum (1). This minimally invasive technique is also used to diagnose benign disorders and lymphoma and to access peripheral nodules and peribronchial abnormalities that might otherwise be inaccessible for safe bronchoscopic sampling.

Thus, a growing number of physicians are seeking to incorporate EBUS-TBNA into their practices. Learning new procedures in the clinical setting, however, promotes learner anxiety, subjects patients to the burden of procedure-related education, and results in a highly variable learning experience (2–5). Time-limited continued medical education programs, although beneficial for familiarizing participants with procedural techniques, are notable for their different training agendas and varied qualities of instruction. These arguments, coupled with an increasing impetus to document procedural competency[1], support the use of objective measures of procedural technical skill in both simulated and clinical environments. The primary aim of using such instruments is to help assure that all learners achieve a benchmark threshold of technical skill and educational outcome (6, 7).

Validity of an assessment instrument or educational measurement refers to the evidence presented to support or refute the meaning or interpretation of data or results obtained with that instrument or measurement (8–11). Thus, to study the validity of a test is to study the accuracy of interpretations and decisions based on the scores derived from that test. The Endobronchial Ultrasound Skills and Tasks Assessment Tool (EBUS-STAT) is specifically designed to score operators' EBUS-TBNA skills. We hypothesized that (*1*) the EBUS-STAT has a high interrater

[1]In the United States this is done in accordance with Accreditation Council for Graduate Medical Education guidelines.

(*Received in original form November 8, 2011; accepted in final form July 2, 2012*)

Author Contributions: M.D. and H.G.C. created the EBUS-STAT assessment tool, designed and executed the study, and prepared the manuscript. J.J.M. and C.R.L. participated in the study design and execution and preparation of the manuscript. K.E.O. participated in the study design and execution, performed the statistical analysis, and participated in preparation of the manuscript.

Correspondence and requests for reprints should be addressed to Mohsen Davoudi, M.D., Division of Pulmonary and Critical Care Medicine, University of California Irvine, Orange, CA 92868. E-mail: mdavoudi@uci.edu

This article has an online supplement, which is accessible from this issue's table of contents at www.atsjournals.org

Am J Respir Crit Care Med Vol 186, Iss. 8, pp 773–779, Oct 15, 2012

Originally Published in Press as DOI: 10.1164/rccm.201111-1968OC on July 26, 2012
Internet address: www.atsjournals.org

reliability, with reproducible results when used by different independent testers; and (*2*) our results would establish evidence for the intended interpretation of operator scores obtained by the EBUS-STAT (8, 10, 12). Thus, we hypothesized that EBUS-STAT scores will accurately distinguish the EBUS-TBNA skill level of beginner, intermediate, and experienced operators, in the appropriate subjects and settings.

METHODS

The Instrument

The EBUS-STAT is a 10-section assessment tool designed to objectively and systematically evaluate the technical skill and relevant knowledge of an operator performing convex-probe (CP) EBUS-guided TBNA. Created as a component of the Bronchoscopy Education Project (13), it can be used alone or in addition to other learning tools, reading materials, and simulation-based educational sessions to document the gradual acquisition of knowledge and skills in learners training to become competent EBUS-TBNA operators.

The EBUS-STAT can be scored while observing an operator perform CP EBUS-TBNA in a patient or simulated environment[2]. Of the 10 items, items 1 to 7 test technical skill, and items 8 to 10 use a 25-image slideshow to test computed tomography (CT) and EBUS image and pattern recognition, anatomic orientation, and correlation. The first seven items are tested in the procedure suite or simulation center, whereas items 8 to 10 can be completed using a computer monitor.

Item 1 tests advancing the EBUS bronchoscope into the airway (through an oral bite-block, laryngeal mask airway, or endotracheal tube); item 2 tests white-light atraumatic navigation of the CP-EBUS bronchoscope in the central airways to each of the lobar orifices; item 3 tests obtaining an artifact-free image and troubleshooting existing artifacts (using basic maneuvers such as more or less inflation of the balloon or better apposition of the CP-EBUS tip); item 4 requires the subject to image five main mediastinal vascular structures (aorta, pulmonary artery trunk, superior vena cava, azygous vein, and left atrium); item 5 requires the subject to show three of the paratracheal/bronchial nodal stations (in cases in which fewer than three stations are visible, the subject can identify the normal location of a station, say the triangle between the trachea, aorta, and pulmonary artery for station 4L); item 6 involves step-by-step sampling of one nodal station (*see* online supplement); and item 7 tests the subject's ability to modify gain, depth, and use Doppler on the EBUS image-processor console.

The Testing Protocol

Three volunteer study centers with established interventional pulmonary programs were chosen based on availability of a volume of EBUS-TBNA cases, such that all assessments could be performed in a 2-day period (University of California Irvine Medical Center [Orange, CA], Lahey Clinic [Burlington, MA], and Mayo Clinic [Rochester, MN]). Study participants were asked to answer a two-item experience questionnaire regarding the approximate number of lifetime flexible and EBUS-TBNA bronchoscopies performed. Based on the number of EBUS-TBNA bronchoscopies, the subjects were stratified into three categories: beginner subjects had performed fewer than 20 EBUS bronchoscopies, intermediate subjects were those with 20 to 50, and experienced subjects had performed more than 50. Subjects with fewer than 100 flexible bronchoscopies were excluded[3]. Subjects were also asked to complete the EBUS Self-Assessment Tool (EBUS-SAT), composed of 10 five-point Likert-style questions pertaining to one's own skills in 10 different EBUS domains (*see* online supplement). This was done to correlate each subject's self-assessment and test score and explore associations between the level of expertise and the accuracy of self-assessment.

Each subject was simultaneously and independently observed and tested by two testers (H.G.C. and M.D.), who did not have access to each other's test sheet. Each subject was given a site-specific identification number, identically printed on the test sheets, experience questionnaire, slideshow answer sheet, and EBUS-SAT. Sheets with matching identification numbers were stapled together; statistical analysis was later performed by an investigator (K.E.O.) who did not know or observe the subjects.

Tests at all three sites were performed on real patients during patients' standard course of care. Observation and testing was only performed if the patients stated that they did not object to two visiting physicians (H.G.C. and M.D.) observing the case. During the procedure, items 1 to 5 and item 7 were scored as the subject performed EBUS-TBNA as part of routine patient care duties: introduction of the CP-EBUS into the airway, central airway examination, ultrasound image obtained, vascular and nodal structures examined, and EBUS-TBNA sampling performed. The subject was simply asked to say out loud explicitly what they were doing, where they were, and what they were seeing on the screen with white-light or ultrasound. Item 6 was tested at the end of the sampling: the subject was asked to modify gain and depth and apply/turn off Doppler to show familiarity with the console. The 25-slide portion of the test (items 8–10) was available on USB flash drives and administered based on convenience, either before or after technical skill testing. It was emphasized before each observation session that the EBUS procedure would be planned and performed as it must for the patient's care. In cases in which the patient needed multiple nodes sampled, subjects who had not been tested were not allowed into the procedure room, to prevent prior exposure bias. At all times, an expert EBUS operator faculty physician was present and supervising the entire procedure, ensuring that all standard-of-care precautions and procedures were observed, protecting the patient's safety and best interest as an absolute priority. Participation in the study was strictly voluntary and anonymous, with no impact on the participants' professional position. The testing protocol was considered exempt from institutional review board review by the University of California Irvine Institutional Review Board Office of Research.

Statistical Methods

The primary aim of the study was to show that the EBUS-STAT has intertester reliability and reproducibility and that EBUS-STAT scores statistically correlate with operator levels of skill, established based on prior experience, thus allowing for accurate interpretations and decisions based on these scores. Intertester agreement was assessed using intraclass correlations to measure agreement between scores given by the two independent testers. Results for all subjects were divided into three groups based on lifetime number of EBUS-TBNA performed: group 1 with fewer than 20, group 2 with 20 to 50, and group 3 with more than 50. Sample size was calculated using nQuery 7.0. With eight subjects per group (total = 24), and assuming a common SD of 30, a one-way analysis of variance would have 80% power to detect at the 5% significance level an effect size of 0.5 (variance of means = 450). The Average Total Score was compared between groups using analysis of variance. *Post hoc* comparisons were conducted to test for pairwise differences between groups with adjustment for multiple comparisons using the Tukey method. The same comparisons were repeated separately for the scores obtained from the technical skills and slideshow sections of EBUS-STAT, items 1 to 7 (subtotal 1), and items 8 to 10 (subtotal 2). Associations between the average total score and number of EBUS-TBNA procedures completed were investigated using linear and nonlinear regression methods. Agreement between self-assessment (based on EBUS-SAT, rated 1–5) and test score (based on EBUS-STAT, scored 0–100) was assessed using Pearson correlations. Internal consistency of the total and subtotal scores was measured by Cronbach α.

RESULTS

Intertester reliability, when tested using intraclass correlations to measure agreement between overall scores given by the two independent testers, was very high ($r = 0.999$, $P < 0.0001$) (Table 1).

[2]Items 4 and 5 consist of the identification of mediastinal anatomy. This cannot be performed using a low-fidelity model, but can be done using high-fidelity simulation. All other items can be tested using a low-fidelity airway inspection and EBUS-TBNA model (14).

[3]Prior studies suggested that a minimum number of flexible bronchoscopies may be necessary to assure competent navigation of the central airway (2, 6). These numbers might not be relevant in the future, as objective measures of technical skill such as the Bronchoscopy Skills and Tasks Assessment Tool are used more regularly to ascertain acquisition of skills (15).

TABLE 1. INTERRATER AGREEMENT

Variables	Intraclass Correlation	*P* Value
Subtotal 1	0.9988	< 0.00005
Subtotal 2	1.00000	N/A
Total	0.9991	< 0.00005

Definition of abbreviation: N/A = not applicable.

Intraclass correlation for total scores for all 24 subjects, along with subtotal 1 (skill test, items 1–7) and subtotal 2 (slideshow, items 8–10). Each subject took the slideshow only once, hence the 1.0 or 100% correlation.

Subjects' test scores were divided into three groups of eight (total n = 24) based on the number of EBUS-TBNA procedures completed: group 1 (n = 8) fewer than 20, group 2 (n = 8) 20 to 50, and group 3 (n = 8) more than 50. This was also done separately for the seven technical items tested at bedside (subtotal 1) and the three computer-based slideshow items (subtotal 2). The mean number of EBUS-TBNA procedures completed before testing for each group were 4.9 (range, 0–15) for group 1, 32.5 (range, 20–45) for group 2, and 154.4 (range, (55–300) for group 3.

Average total scores for the three groups were 31.1 (SE, 3.75) for group 1, 74.9 (SE, 2.57) for group 2, and 93.6 (SE, 2.32) for group 3 ($F_{2,21} = 118.6$, $P < 0.0001$). *Post hoc* tests showed each group differed significantly from the other, with $P < 0.001$ after adjustment for multiple comparisons (Figures 1A–1C).

Average subtotal 1 scores (out of 70) for the three groups were 13.75 (SE, 3.75) for group 1, 54.50 (SE, 2.57) for group 2, and 68.38 (SE, 2.32) for group 3 ($F_{2,21} = 162.91$, $P < 0.00005$). *Post hoc* (Tukey) tests showed each group differed significantly from the other with $P < 0.001$ after adjustment for multiple comparisons. Average subtotal 2 scores (out of 30) for the three groups, respectively, were 17.38 (SE, 3.75) for group 1, 20.38 (SE, 2.57) for group 2, and 25.25 (SE, 2.32) for group 3 ($F_{2,21} = 10.65$, $P = 0.0006$). *Post hoc* (Tukey) tests showed that group 3 differed significantly from group 1 ($P = 0.0005$) and group 2 ($P = 0.026$); the difference of 17.38 and 20.38 between groups 1 and 2 did not reach 95% significance ($P = 0.21$). Internal consistency as measured by Cronbach α was 0.85, 0.77, and 0.86 for subtotal 1, subtotal 2, and total score, respectively.

Using simple regression methods, we further explored the relationship between average total score and lifetime number of EBUS-TBNAs performed, as a continuous variable. This relationship proved to be nonlinear, with the logarithm of the number of lifetime EBUS-TBNAs significantly predicting the average total score, with $r^2 = 0.94$ ($P < 0.001$). When groups 1 and 2 alone (< 50 total EBUS-TBNAs) are investigated, there is a significant linear association between number of lifetime EBUS-TBNAs and average total score ($P < 0.001$, $r = 0.88$). Completion of more than 50 EBUS-TBNAs, however, contributed little to the improvement in test score (Figures 2A–2C).

Exploring the best multivariate predictive model for total score, we found that the prior number of EBUS-TBNA bronchoscopies as a categorical variable (stratified into groups 1, 2, or 3) was the best predictor of average total score, with $r^2 = 0.92$. It was observed that the groups are basically nonoverlapping with respect to total score (Figures 1A–1C). Over the entire sample, the relationship is not linear where the total score is correlated with the logarithm of the number of EBUS-TBNA procedures (Figures 2A–2C). It seems that once a practitioner has completed more than 50 EBUS-TBNAs, they test above 80%, and three-fourths test above 90%, so there may be little room for improvement in EBUS-STAT scores with more experience. For those in groups 1 and 2 (who have completed fewer than 50 EBUS-TBNAs), there was a linear relationship between the number of EBUS-TBNA procedures and total score.

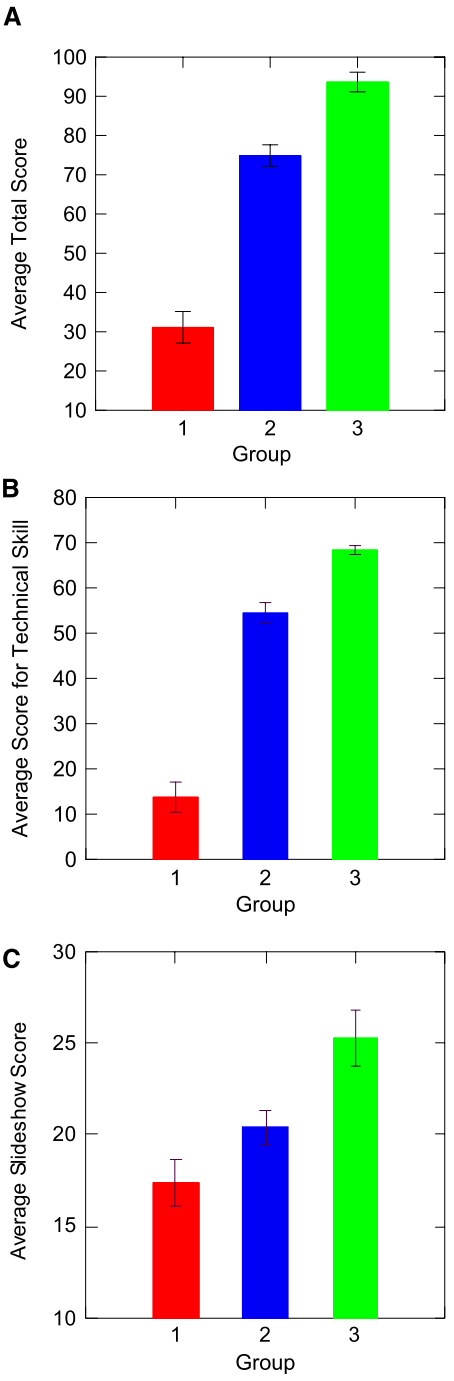

Figure 1. (*A*) Average total scores for three groups (maximum score = 100). Average total scores (out of 100) for groups 1–3, respectively, were 31.1 (SE, 3.75; range, 18–46), 74.9 (SE, 2.57; range, 58–82), 93.6 (SE, 2.32; range, 82–99); ($F_{2,21} = 118.6$, $P < 0.0001$). *Post hoc* (Tukey) tests showed each group differed significantly from the other with $P < 0.001$ after adjustment for multiple comparisons. (*B*) Average subtotal 1 scores for three groups (maximum score = 70). Average subtotal 1 scores (out of 70) for groups 1–3, respectively, were 13.75 (SE, 3.75), 54.50 (SE, 2.57), 68.38 (SE, 2.32); ($F_{2,21} = 162.91$, $P < 0.00005$). *Post hoc* (Tukey) tests showed each group differed significantly from the other with $P < 0.001$ after adjustment for multiple comparisons. (*C*) Average subtotal 2 scores for three groups (maximum score = 30). Average subtotal 2 scores (out of 30) for groups 1–3, respectively, were 17.38 (SE, 3.75), 20.38 (SE, 2.57), 25.25 (SE, 2.32); ($F_{2,21} = 10.65$, $P = 0.0006$). *Post hoc* (Tukey) tests showed group 1 < group 3 ($P = 0.0005$), group 2 < group 3 ($P = 0.026$), group 1 < group 2 did not reach 95% significance ($P = 0.21$).

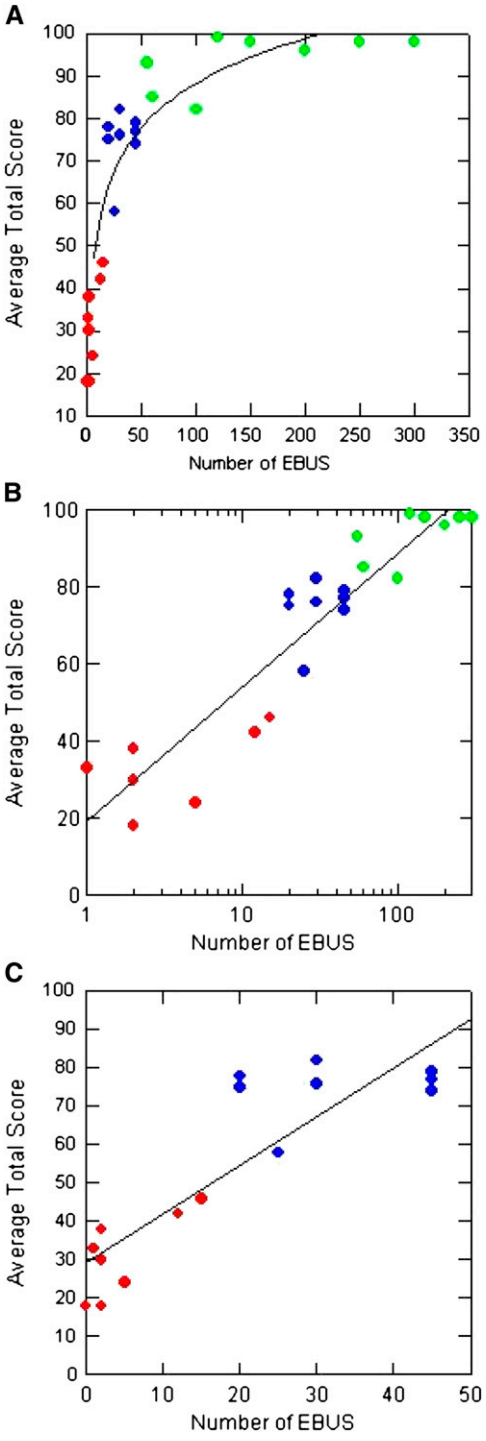

Figure 2. Linear and log-linear plots demonstrating association of three groups' average total scores with number of endobronchial ultrasound-guided transbronchial needle aspirations (EBUS-TBNAs) performed. (*A*) When plotted using a linear scale for both axes, the number of EBUS-TBNAs in groups 1 and 2 shows a solid linear association of average total score with number of EBUS for groups 1 and 2, which becomes less clear for group 3 (i.e., after 50 or more EBUS-TBNAs). This is suggestive of a log-linear distribution. (*B*) When plotted in log-linear fashion (plotting the number of EBUS on the X-axis in logarithmic scale) we obtain a perfect straight line for groups 1 through 3. (*C*) Alternately, by excluding group 3 (subjects with 50 or more EBUS-TBNAs), the linear plot of groups 1 and 2 also renders a perfect straight line.

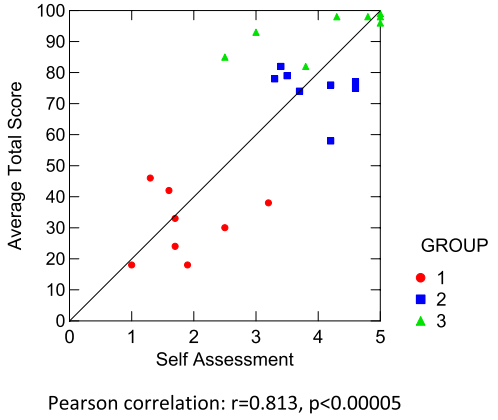

Figure 3. Association between test scores and self-assessment (Endobronchial Ultrasound Skills and Tasks Assessment Tool vs. Endobronchial Ultrasound Self-Assessment Tool). The *diagonal line* indicates what would be perfect agreement.

Self-assessment of ability by subjects (EBUS-SAT) corresponded closely to actual test scores, with $r^2 = 0.81$ and $P < 0.001$ (Figure 3). In general, more subjects overestimated their ability (relative to total score) than underestimated, with the exception of group 3, in whom there was a tendency to underestimate ability. Although one-half of the subjects in groups 1 and 2 rated their ability above their test level (four of eight in each of groups 1 and 2), subjects in group 3 were more likely to underestimate their ability.

Initial power calculations, using nQuery 7.0, had assumed a common SD = 30 and a difference between means characterized by a variance of means for the three groups of 450 (effect size = 0.5). Hence, to detect this effect with 80% power at a 0.05 significance level a sample of eight subjects per group (total = 24) was required. A *post hoc* power calculation was to ascertain that with the real SD and variance, we had achieved our target 80% power. *Post hoc* power calculations using actual data and variances demonstrated a power of 98% to detect differences.

DISCUSSION

The late Christine McGuire, a leader of assessment in medical education, wrote, "Evaluation is probably the most logical field in the world and if you use a little bit of logic, it just fits together and jumps at you . . . It's very common sense" (16). An established common-sense tenet of competency-oriented medical education has been that all competencies must be teachable, learnable, and measurable (17). This study set out to demonstrate the ability of the EBUS-STAT to provide objective, formative, and summative assessments of EBUS-TBNA–related knowledge and skills. This would require that we establish evidence for the accurate interpretation of scores derived from this assessment tool when used in appropriate subjects and settings.

The necessity for validation may be a foregone conclusion, but the design of any test of clinical skills "should always include

[4]Validity, or "a test measuring what it is supposed to measure (18, 19)" has been traditionally divided into various elements, including *construct*, *content*, *criterion* (divided into *concurrent* and *predictive*), and *face validity* (the latter eschewed by most contemporary educational assessment researchers on the grounds that appearance is not scientific evidence, and appearance of validity must not be equated with validity (11, 20). Downing, Haladyna, and others, however, believe this view of validity to be dated (10, 21, 22) and that "validity is a unitary concept, which requires multiple sources of scientific evidence to support or refute the meaning associated with assessment data" (10, 22). Hence, in this framework, all validity leads to construct validity.

test validation studies during the early stages" (20, 21). Downing states that, "Validity[4] is the *sine qua non* of assessment, as without evidence of validity, assessments in medical education have little or no intrinsic meaning" (8), and refers to validity as "the single most important topic in testing."

One respected method of test validation is the "use of clinical validity groups to show discriminant validity" (23, 24). Roid states that, "One simple index of construct validity for most ability tests is the sensitivity of items to the developmental trend" (23). This is precisely the premise that we pursued. Our results demonstrate that our "ability test" can objectively discriminate between operators at three levels of experience from novice to expert.

A prerequisite of validity is "fidelity to the criterion," which has been defined as "some *validity-type* relationship between scores or ratings on the assessment and the ultimate '*criterion*' variable in real life" (18, 19, 25). Essentially, fidelity to the criterion requires a mathematical relationship between the score on the valid test to the real-life variable that the test purports to measure. As demonstrated by the results, in the case of the EBUS-STAT, scores on the assessment tool had a highly significant relationship to the real-life variable of EBUS-TBNA operator experience and expertise.

The other major requirement of a test is reliability (12, 26–28), defined as the proportion of reproducible data to random noise recorded by the assessment instrument. A reliability of 1.0 implies the data are 100% objectively representative of the subject matter being measured and will be reproducible exactly, again and again; a reliability of 0.0 means the data are all noise, with no objective bearing to the facts being recorded. The EBUS-STAT had a reliability of 0.9991, and the bedside section (items 1–7) has a reliability of 0.9988, attesting to a high level of reproducibility, obviously a desirable trait for a test of procedural skills.

Another important characteristic that renders a test pragmatically more useful is "examiner and examinee friendliness" (23). The EBUS-STAT was designed with such user friendliness in mind. Scoring follows clear and reproducible instructions and definitions; the examinee and examiner know precisely what is expected and being tested in each of the bedside items, which can be tested without any interruption to the course of patient care. The computer-based slideshow is self-explanatory and readily completed outside the patient-care setting.

Comparison of the subjects' self-assessment using EBUS-SAT with actual EBUS-STAT scores, although demonstrating a high degree of correlation ($r = 0.81$, $P < 0.001$), revealed that more subjects overestimate their ability than underestimate, with the exception of the expert group, in which there was a tendency to underestimate ability. These findings confirmed two well-known phenomena: (*1*) most operators overestimate their skills, and (*2*) less experienced operators have a stronger tendency to do so (29–33).

One limitation of this study is that both testers were closely familiar with the assessment tool; hence, the high interrater reliability may not be universal. Yet, this issue is true for all checklist and global rating scale tests. It is precisely why testers need to become familiar with their testing instruments and why the assessment process must be calibrated to prevent each training program from developing groups of "faculty 'hawks' and 'doves,'" each known for extra stringency or leniency[5] (18, 19, 34–37). Another limitation is that when EBUS STAT is used repeatedly to plot the learner's progression along the learning curve, the slideshow section (items 8–10) may become irrelevant once the learner masters it. Thus, the fact that the difference between average subtotal 2 scores for groups 1 and 2 did not reach statistical significance (17.38 vs. 20.38, $P = 0.21$) is inconsequential, because this section of the test was never meant to be used separately *per se*. The reason to report the subtotal scores was to show that the significant difference between the total scores of the three groups is primarily due to subtotal 1 differences (i.e., differences in the bedside technical skills and nonslideshow part of the test). Hence, when the test is administered repeatedly to plot a learner's progress, the skills section (i.e., subtotal 1, items 1–7) will still be useful when used separately. Finally, because testing was done in patients, it is possible that individual case difficulty could have randomly biased results (38). This is unlikely because (*1*) a conscious effort was made to test subjects from all three experience levels on the same patients, and (*2*) it was clearly observed that a novice would not perform better on the test when tested on an "easy" patient, nor would experienced operators do worse on the "difficult" patients.

A result that warrants consideration is our observation that experience beyond what is attained by completion of more than 50 to 100 EBUS procedures contributed little to the improvement in test score (Figures 2A–2C). This obvious ceiling effect was indeed intentional and one of the primary considerations in the design of EBUS-STAT. This assessment tool was designed for mastery testing (39–43)[6], which is a type of criterion-referenced testing (19, 34, 44, 45). Criterion-referenced testing (as opposed to norm-referenced testing), which is closely associated with competency-based education and assessment (19, 34), explores how much (or what proportion) of some specific content of knowledge and skills the learners know or can do. In mastery training and testing (39–43), the assessments are devised such that they must be completed nearly perfectly by almost all learners. For a mastery test, the expected score is 100% correct. Bronchoscopic procedures such as EBUS-TBNA are ideal for mastery training and testing, as every operator must master each of the constituent elements of a safe and effective procedure to achieve competency. The variable that distinguishes different learners is the slope of the curve (i.e., time required by each learner to reach this educational goal) (19, 39–43, 46).

Learners and instructors are encouraged to see the mastery model of training and assessment as an expression of a dynamic testing paradigm (47), whereby training and assessment merge into one; students learn to see the assessment process and score as their friend and not their foe. Instructors can use assessment tools to plot each learner's acquisition of knowledge and skills, identify strengths and weaknesses, and design learner-centric remedial training, individualizing sessions to focus on each learner's deficiencies.

This begs the question: is there no difference in expertise after 50 or 100 EBUS-TBNA procedures? The answer is an emphatic *No*. Acquisition of fine nuances of procedural excellence continues for years after competency benchmarks have been surpassed. Much in the same way that the Objective Structured Clinical Examination was not meant to distinguish between masters of bedside medicine, the EBUS-STAT was not designed to measure nuances of excellence but to ensure

[5]This information can be covered during faculty development programs, such as *train the trainers*.

[6]Researchers have proposed that mastery testing is ideally suited for the ultimate goal of using the *mastery model of training and assessment* in the medical field, especially as it pertains to clinical and procedural competencies (16, 38–42).

attainment of the fundamental knowledge and skill elements required for a safe and effective EBUS-TBNA (48, 49).

Conclusions

The evolving paradigm of mastery training in procedural medical education necessitates the development and validation of reliable assessment tools to objectively measure the acquisition of technical skill, documenting each learner's progress along the learning curve from novice to competent practitioner. We demonstrate that the EBUS-STAT can be used to reliably and objectively score and classify EBUS-TBNA operators from novice to expert. It has a user-friendly and logical structure and may be administered at the bedside or on a combination of low- and high-fidelity simulation platforms. Its use to assess and document the acquisition of knowledge and skill among learners is a step toward the goal of mastery training in EBUS-TBNA.

Author disclosures are available with the text of this article at www.atsjournals.org.

Acknowledgment: The authors thank Dr. Steven Downing. His scholarship has guided the premise of their work, and he has graciously assisted them with the writing and revisions of the manuscript.

References

1. Colt HG, Davoudi M, Murgu SD. Scientific evidence and principles for the use of endobronchial ultrasound and transbronchial needle aspiration. *Expert Rev Med Devices* 2011;8:493–513.
2. Pastis NJ, Nietert PJ, Silvestri GA. Variation in training for interventional pulmonary procedures among us pulmonary/critical care fellowships: A survey of fellowship directors. *Chest* 2005;127:1614–1621.
3. Unroe MA, Shofer SL, Wahidi MM. Training for endobronchial ultrasound: methods for proper training in new bronchoscopic techniques. *Curr Opin Pulm Med* 2010;16:295–300.
4. Davoudi M, Colt HG. Bronchoscopy simulation: a brief review. *Adv Health Sci Educ Theory Pract* 2009;14:287–296.
5. Stather DR, Lamb CR, Tremblay A. Simulation in flexible bronchoscopy and endobronchial ultrasound: a review. *J Bronchology Interv Pulmonol* 2011;18:247–256.
6. Lamb CR, Feller-Kopman D, Ernst A, Simoff MJ, Sterman DH, Wahidi MM, Kovitz KL. An approach to interventional pulmonary fellowship training. *Chest* 2010;137:195–199.
7. McGaghie WC, Siddall VJ, Mazmanian PE, Myers J. Lessons for continuing medical education from simulation research in undergraduate and graduate medical education: effectiveness of continuing medical education: American College of Chest Physicians evidence-based educational guidelines. *Chest* 2009;135:62–68.
8. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ* 2003;37:830–837.
9. Messick S. Validity. In: Linn RL, editor. Educational measurement, 3rd ed. New York, NY: American Council on Education & Macmillan; 1989. pp. 13–104.
10. Downing SM, Haladyna TM. Validity and its threats. In: Downing SM, Yudkowsky R, editors. Assessment in health professions education. New York, NY: Routledge; 2009. pp. 21–55.
11. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 1999.
12. Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ* 2004;38:1006–1012.
13. Bronchoscopy International. The Bronchoscopy Education Project (Part II): Endobronchial Ultrasound Bronchoscopy Competency Program; 2011 [accessed 2011 Sept 25]. Available from: http://www.aabronchology.org/PDF/bronch_course/BEP%20part%20II%20EBUS%20Training%20Manual.pdf
14. Goldberg R, Colt HG, Davoudi M, Cherrison L. Realistic and affordable lo-fidelity model for learning bronchoscopic transbronchial needle aspiration. *Surg Endosc* 2009;23:2047–2052.
15. Wahidi MM, Silvestri GA, Coakley RD, Ferguson JS, Shepherd RW, Moses L, Conforti J, Que LG, Anstrom KJ, McGuire F, *et al*. A prospective multi-center study of competency metrics and educational interventions in the learning of bronchoscopy among starting pulmonary fellows. *Chest* 2010;137:1040–1049.
16. Harris IB, Simpson D. Christine McGuire: at the heart of the maverick measurement maven. *Adv Health Sci Educ Theory Pract* 2005;10:65–80.
17. Brasel KJ, Bragg D, Simpson DE, Weigelt JA. Meeting the Accreditation Council for Graduate Medical Education competencies using established residency training program assessment tools. *Am J Surg* 2004;188:9–12.
18. Downing SM. Twelve steps for effective test development. In: Downing SM, Haladyna TM, editors. Handbook of test development. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers; 2006. pp. 3–25.
19. Downing SM, Yudkowsky R. Introduction to assessment in the health professions. In: Downing SM, Yudkowsky R, editors. Assessment in health professions education. New York, NY: Routledge; 2009. pp. 1–20.
20. Downing SM. Face validity of assessments: faith-based interpretations or evidence-based science? *Med Educ* 2006;40:7–8.
21. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006;119:e7–e16.
22. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ* 2004;38:327–333.
23. Roid GH. Designing ability tests. In: Downing SM, Haladyna TM, editors. Handbook of test development. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers; 2006. pp. 527–542.
24. Kane M. Content related validity evidence in test development. In: Downing SM, Haladyna TM, editors. Handbook of test development. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers; 2006. pp. 3–25.
25. Yudkowsky R. Performance tests. In: Downing SM, Yudkowsky R, editors. Assessment in health professions education. New York, NY: Routledge; 2009. pp. 217–243.
26. Axelson RD, Kreiter CD. Reliability. In: Downing SM, Yudkowsky R, editors. Assessment in health professions education. New York, NY: Routledge; 2009. pp. 57–73.
27. Stemler SE. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation* 9(4). 2004 [accessed 2011 Sept 25]. Available from: http://PAREonline.net/getvn.asp?v=9&n=4
28. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174.
29. Davis DA, Mazmanian PE, Fordis M, Harrison VR, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA* 2006;296:1094–1102.
30. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J Pers Soc Psychol* 1999;77:1121–1134.
31. Ehrlinger J, Johnson K, Banner M, Dunning D, Kruger J. Why the unskilled are unaware: further explorations of (absent) self-insight among the incompetent. *Organ Behav Hum Decis Process* 2008;105:98–121.
32. Gordon MJ. A review of the validity and accuracy of self-assessments in health professions training. *Acad Med* 1991;66:762–769.
33. Regehr G, Hoges B, Tiberius R, Lofchy J. Measuring self-assessment skills: an innovative relative ranking model. *Acad Med* 1996;71:S52–S54.
34. Yudkowsky R, Downing SM, Tekian A. Standard setting. In: Downing SM, Yudkowsky R, editors. Assessment in health professions education. New York, NY: Routledge; 2009. pp. 119–148.
35. Cizek CJ. Standard setting. In: Downing SM, Haladyna TM, editors. Handbook of test development. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers; 2006. pp. 225–258.
36. Norcini JJ. Setting standards on educational tests. *Med Educ* 2003;37:464–469.
37. Kane M. Validating the performance standards associated with passing scores. *Rev Educ Res* 1994;64:425–461.

38. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med* 2003;15:270–292.
39. Wayne DB, Butter J, Siddall VJ, Fudala MJ, Wade LD, Feinglass J, McGaghie WC. Mastery learning of advanced cardiac life support skills by internal medicine residents using simulation technology and deliberate practice. *J Gen Intern Med* 2006;21:251–256.
40. Downing SM, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teach Learn Med* 2006;18:50–57.
41. Luecht RM. Designing tests for pass-fail decisions using item response theory. In: Downing SM, Haladyna TM, editors. Handbook of test development. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers; 2006. pp. 575–596.
42. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Acad Med* 2004;79:S70–S81.
43. Zendejas B, Cook DA, Bingener J, Huebner M, Dunn WF, Sarr MG, Farley DR. Simulation-based mastery learning improves patient outcomes in laparoscopic inguinal hernia repair: a randomized controlled trial. *Ann Surg* 2011;254:502–511.
44. Kane MT. The role of reliability in criterion-referenced tests. *J Educ Meas* 1986;23:221–224.
45. Popham WJ, Husek TR. Implications of criterion-referenced measurement. *J Educ Meas* 1969;6:1–9.
46. Wolcox RR. A note on the length and passing score of a mastery test. *J Educ Stat* 1976;1:359–364.
47. Grigorenko EL, Sternberg RJ. Dynamic testing. *Psychol Bull* 1998;124:75–111.
48. Hodges B, Regehr G, McNaugton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Acad Med* 1999;74:1129–1134.
49. Newble D, Dawson B, Dauphinee D, Page G, Macdonald M, Swanson D, Mulholland H, Thomson A, van der Vleuten C. Guidelines for assessing clinical competence. *Teach Learn Med* 1994;6:213–220.