Development and Validation of a Theoretical Test in Endosonography for Pulmonary Diseases

Mona M. Savran^a Paul Frost Clementsen^b Jouke T. Annema^c
Valentina Minddal^b Klaus R. Larsen^d Yoon Soo Park^e Lars Konge^a

^aCentre for Clinical Education, University of Copenhagen and the Capital Region of Denmark, Copenhagen, and

^bDepartment of Pulmonology, Gentofte Hospital, University of Copenhagen, Hellerup, Denmark;

^cDepartment of Pulmonology, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands;

^dDepartment of Pulmonology, Bispebjerg Hospital, University of Copenhagen, Copenhagen, Denmark;

^eDepartment of Medical Education, University of Illinois Chicago, Chicago, Ill., USA

Key Words

Theoretical test · Validity testing · Endobronchial ultrasonography · Esophageal ultrasonography

Abstract

Background: Theoretical testing provides the necessary foundation to perform technical skills. Additionally, testing improves the retention of knowledge. **Objectives:** The aims of this study were to develop a multiple-choice test in endosonography for pulmonary diseases and to gather validity evidence for this test. **Methods:** Initially, 78 questions were constructed after informal conversational interviews with 4 international experts in endosonography. The clarity and content validity of the questions were tested using a Delphi-like approach. Construct validity was explored by administering the test to 3 groups with different levels of endosonography experience: 27 medical students, 18 respiratory physicians with limited endosonography experience, and 14 experts in endosonography. **Results:** Two Delphi iterations reduced the test to 52 questions. After item analysis, the final test consisted of 46 questions with a mean item discrimina-

tion of 0.47 and a mean item difficulty of 0.63. The internal consistency reliability was calculated at 0.91. The 3 groups performed significantly differently (ANOVA: $p < 0.001$), and post hoc tests were significant. The experts performed significantly more consistently than the novices ($p = 0.037$) and the intermediates ($p < 0.001$). **Conclusions:** This study provides a theoretical test in endosonography consisting of multiple-choice questions. Validity evidence was gathered, and the test demonstrated content and construct validity.

© 2014 S. Karger AG, Basel

Introduction

Accurate staging of non-small cell lung cancer is pivotal in determining therapeutic strategies, and the evaluation of mediastinal lymph nodes through tissue verification is a key component [1, 2]. Staging can be performed in a minimally invasive manner via endosonography including endobronchial ultrasonography (EBUS) and esophageal ultrasonography (EUS) with real-time aspiration of the lymph nodes [3–5]. Unlike surgical staging

techniques such as mediastinoscopy, EBUS and EUS are minimally invasive and cost-effective. Current guidelines from American College of Chest Physicians (ACCP), the European Society of Thoracic Surgeons (ESTS), and the European Society for Medical Oncology (ESMO) regarding lung cancer staging state that, in cases where mediastinal nodal tissue staging is indicated, EBUS and EUS are the staging procedures of choice before opting for any surgical techniques [1, 2, 6].

Both EBUS and EUS are challenging procedures dependent on skill acquisition [7–10]. The acquisition of cognitive skills provides trainees with the necessary foundation to properly perform technical skills [11]. Predominantly, the purpose of testing is to assess theoretical knowledge (i.e. cognitive skills), but testing also encompasses the opportunity to increase the later retention of information and thereby promote learning; this is also known as the ‘testing effect’ [12, 13]. A supporting argument for testing theoretical knowledge is that testing is superior to studying in terms of enhancing long-term retention [14]. Other major purposes of testing are to communicate what are viewed as paramount goals in the curriculum and to motivate to self-study [15].

Since tests have a fundamental impact on learning, the following aspects must be thoroughly considered in the development of test questions: type, format, content, validity, reliability, educational impact, cost-effectiveness, and acceptability [15–17]. Written tests are more cost-effective and reliable than other assessment types [18]. Several formats of written tests exist, including multiple-choice questions (MCQs) (e.g. one-best-answer questions, true-false questions, and extended matching questions) [15–18]. MCQs have several advantages that make them suitable for testing cognitive skills: encompassment of a wide range of the curriculum, short answering and scoring times, reuse on future tests, and high reliability [16–18]. Above all, MCQs are not only efficient in measuring information learned by rote (i.e. ‘recall questions’) but they can also ensure the use of knowledge [15].

Currently, such procedure-relevant knowledge testing does not exist in the field of endosonography, although several articles emphasize the importance of learning and implementing both EBUS and EUS in the diagnostic methodology of cancer including non-small cell lung cancer [7, 8, 19, 20]. Therefore, the aims of this study were to develop a theoretical test consisting of MCQs dealing with endosonography (EBUS and EUS), and to gather the necessary validity evidence for the test.

Material and Methods

This study was comprised of two parts: the development of MCQs and gathering of validity evidence (fig. 1).

Expert Interviews (Step 1)

M.M.S. completed informal conversational interviews with 4 international experts in endosonography (P.F.C., J.T.A., V.M., and K.R.L.). The experts represented 3 different centers in 2 European countries (Denmark and The Netherlands). The interviews were based on the question: ‘What do you view as relevant endosonography knowledge in relation to mediastinal assessment?’. Written notes were made during the interviews.

Development of the MCQs (Step 2)

Based on the information gathered from the 4 interviews, MCQs were constructed according to a manual by Case and Swanson [15], advice concerning written tests by Downing [17], and a taxonomy of 31 multiple-choice item-writing guidelines by Haladyna [21]. The selected multiple-choice-item format was one-best-answer questions beginning with a stem which contained information necessary for answering the question (e.g. a clinical vignette) and a lead-in question, and at the bottom a list of choices (a, b, and c) was presented including both the correct answer and two misleading options. Homogeneity in content existed in the 3 choices, but the 2 misleading answers were less correct compared to the most correct choice. Research has shown that 3 choices in an item are ample; more than 2 misleading options are likely to be redundant and indiscriminate [21].

Validity Testing

Inspired by several articles, the setting for gathering validity evidence for the MCQs included content and construct validity testing (steps 3 and 4 in fig. 1) [16, 22–25].

Content Validity (Step 3)

The developed MCQs were distributed to the 4 experts. Using expert judgement in a Delphi-like process, the questions were rated with regard to their relevance until unanimity was reached. A rating scale of 1 (completely irrelevant) to 5 (extremely relevant) was applied to evaluate the test items. After each round, the ratings of the panel were collated. Questions rated ‘1’ by one or more members of the panel were excluded and questions with a mean rating of 4 or less were reevaluated until unanimity was reached. The experts were also instructed to comment on the wording in order to achieve clarity of the items.

Construct Validity (Step 4)

The purpose of this step was to examine whether the MCQ test scores could discriminate between 3 groups consisting of novices, intermediates, and experts representing no, limited, and high endosonography experience. The group of novices was defined as medical students in their seventh semester of medical school who had lately passed their exam in pulmonology. The group of intermediates consisted of respiratory physicians who had signed up for an endosonography course involving both EBUS and EUS. They were tested before the beginning of the endosonography course. The experts were defined as an international group of respiratory physicians selected due to their profound theoretical, technical, and clinical knowledge in the field of endosonography. In all 3 lev-

Fig. 1. The 4 steps in the development of an MCQ test and validity evidence gathering. Data from this study is shown on the right side.

els, no support (e.g. handbooks or the Internet) was allowed during the test and the test takers did not adhere to any time limit. Answering the MCQ test was voluntary, and according to the local ethics committee approval was not required.

Statistical Analysis

Item analysis was performed to identify malfunctioning test items before further analysis. Questions with an item discrimination index (point-biserial) below 0.2 were discarded while carefully considering the distribution of the remaining items to match the test content. The remaining questions were classified as level I to level IV depending on item difficulty [26]. Internal consistency reliability was calculated using Cronbach's α . The MCQ scores of the 3 groups were compared using one-way ANOVA and Bonferroni's correction for multiple comparisons. Levene's test was used to compare the consistency of the performance in the 3 groups. Statistical analysis was performed using a statistical software package (PASW, version 19.0; SPSS Inc., Chicago, Ill., USA). $p < 0.05$ was considered statistically significant.

Results

Content Validity (Step 3)

A total of 78 MCQs were developed. In the first Delphi iteration, 15 questions were discarded and in the second and final Delphi iteration 11 questions were excluded. Therefore, the MCQ test version, which was handed out to the 3 groups, contained 52 questions; 14 of these had a picture in the stem. In both Delphi iterations the response rate was 100%.

Construct Validity (Step 4)

Six of the questions (12%) had an item discrimination index below 0.2. These questions were excluded from the MCQ test. The mean item discrimination index of the final 46 questions was 0.47 (SD 0.17), and the internal con-

Table 1. The 3 group performances

	Total, n	Mean score	SD	95% CI for the mean		Minimum score	Maximum score
				lower limit	upper limit		
Novices	27	18.6	3.2	17.3	19.8	13.0	25.0
Intermediates	18	31.3	5.9	28.4	34.3	20.0	38.0
Experts	14	40.1	1.9	39.0	41.2	37.0	43.0

Fig. 2. Box-and-whiskers plot illustrating the MCQ test scores of the 3 groups. The lower and upper whiskers show the minimum and maximum scores. The median scores are also represented. The 3 groups performed significantly differently ($p < 0.001$).

sistency reliability was high (Cronbach's $\alpha = 0.91$). The mean item difficulty was 0.63, and the items were classified as: level I (middle difficulty) 70%, level II (easy) 13%, level III (difficult) 13%, and level IV (extremely difficult or easy) 4%. Table 1 demonstrates the performances of the 3 groups. The groups performed significantly differently (F statistic = 142.5; $p < 0.001$), and post hoc tests showed that the intermediate group performed significantly better than the novices ($p < 0.001$) and significantly worse than the experts ($p < 0.001$). The mean difference between the intermediates and the novices was 12.7 and between the intermediates and the experts it was -8.7 (fig. 2). The experts performed significantly more consistently than the novices ($p = 0.037$) and the intermediates ($p < 0.001$).

Discussion

The current study demonstrated that the 4 steps shown in figure 1 could be used to develop a theoretical test and gather validity evidence regarding endosonography for pulmonary diseases.

In step 1, the open-ended format of the interview allowed increased depth and relevance and the content was not prescheduled and therefore was only influenced by the single interviewee's field of expertise [27]. This conversational style is regarded as optimal when handling large-scale, comprehensive knowledge. However, it can also be unstructured and time consuming if the interviewees are too dissimilar regarding their knowledge area. In this study, it was considered that choosing 4 respiratory physicians with great theoretical and technical proficiency eliminated these disadvantages.

In step 2, the test was developed according to the guidelines for constructing MCQs. A major priority in the development of the test was to construct items with a high cognitive demand that endorsed the application of endosonography knowledge in a clinical setting instead of simply testing the test takers' recall abilities. This was accomplished by including clinical vignettes in the stem [15].

In step 3, a Delphi-like method was applied to the experts' rating process. This group communication strategy is anonymous and iterative, yielding the sum of experts' opinions on a specific knowledge area [25]. In our version of the Delphi, we did not convey the experts' responses from the first iteration to the second iteration; only the questions they did not agree on the first time were returned to prevent unnecessary influence from a persuasive expert to another, whereby potential subject bias was avoided [28].

In step 4, forty-six items with an item discrimination index above 0.2 were accepted because these were optimal in portraying the aptitudes of the items to discriminate between high-scoring and low-scoring test takers [26] or,

as Haladyna [29] states: 'Those possessing more of the trait should do better in the items constituting the test than those possessing less of that trait.' The final test achieved good item discrimination with a mean of 0.47 (SD = 0.17).

Apart from a high item discrimination, the most informative items also possess a medium item difficulty [26]. The calculated item difficulty mean of 0.63 was appropriate. Our findings regarding the distribution of the items in the 4 levels were also advantageous since most of the items (70%) were placed in level I. This level is characterized by possession of the best item statistics, and it is advised to include these items in the test [26]. Only 13% of the items were placed in level II. These items are considered easier than items in level I; acceptance of these items was based on their item discrimination index which was above 0.2. Both items in level III (difficult) and IV (extremely difficult or easy) are regarded as inapt in educational measurement; nonetheless, we decided to include the items in both levels since they can increase the content validity of the test scores [26].

Performances significantly differed among the 3 groups and the differences were also significant when comparing the groups with each other. This showed a good discriminative ability. Significant findings showed that the experts were not only the highest scoring group compared to the novices and intermediates but they were also more consistent than the 2 other groups. This outcome is coherent with the Fitts and Posner 3-stage model [30]. The model presents 3 stages of learning: the cognitive stage (first stage), the associative stage (second stage), and the autonomous stage (third stage). According to this model, the experts in our research belonged to the autonomous stage because their performance was less variable and much better than those of the medical students and the intermediates, whose performances were characterized by errors and inconsistency.

In certification and other high-stakes assessments, the desired value of internal consistency reliability should be above 0.8 [31], making our value of 0.91 very satisfactory.

Limitations existed in our study. Although using experts' judgments complies with the definition of content validity, the selection of experts could have distorted the content of the expert interviews in step 1, the item ratings in step 3, and the test answering in step 4, resulting in an inadequate point of view, which would yield selection bias [23, 28]. There is no gold standard for classification as experts or guidelines describing how to select experts and how many to select [28, 32]. Streiner and Norman

[32] recommend 3–10 people who are known by the test developer as experts. On the one hand, with this approach these authors state that the test developer will have access to the most recent knowledge in a specific field, but on the other hand they acknowledge the risk of a skewed expert panel [32]. In order to lessen this bias, we chose 4 physicians with great theoretical, technical, and clinical proficiency in endosonography from 3 different centers in 2 countries for steps 1 and 3 to achieve an international perspective in the MCQ test and in step 4 we used 14 international physicians with similar proficiency profiles (fig. 1). Furthermore, since subjectivity was present in the rating process, the reliability of the experts' opinions could be limited [28]. It cannot be excluded that another panel would rate the items differently.

Another limitation of this study concerns the impact of the constructed MCQ test. As mentioned, the aim of theoretical tests is not solely the assessment of knowledge; the integral mnemonic advantage known as the 'testing effect' is also attributed to testing [12, 13]. In the present study, we did not test this effect and therefore it remains to be studied whether this test acts beyond simply being a measurement tool and accomplishes retention of knowledge.

In medical education, knowledge testing is not ample in giving physicians the required competence, and other aspects must be added to accomplish full competency [33]. At the same time, diverging arguments exist regarding the correlation between cognitive and technical domains as parts of the competence notion. The advantage of gained cognitive skills for technical skills was demonstrated in a randomized study [11]. Nonetheless, another study showed a poor relationship between obtained results from a written test followed by a technical examination on the same matter [34]. Thus, the latter study concluded that no dependency exists between cognitive and technical skills. In light of these two studies, we believe that a multimodal assessment strategy for respiratory physicians is desirable for gaining clinical competence. The developed MCQ test can be considered the first step towards achieving the desired competence in endosonography. Technical skill elements can be tested in clinical practice by using CUSUM analysis, virtual-reality simulators, or specific assessment tools for EBUS and EUS such as EBUS-STAT (EBUS skills and tasks assessment tool) and EUSAT (EUS assessment tool) [35–38].

In conclusion, the major role of endosonography (EBUS and EUS) for lung cancer staging implies the major importance of achieving the knowledge base in this

area. This study yielded a theoretical test in endosonography consisting of MCQs based on experts' knowledge in endosonography. Content and construct validity were established and we recommend implementing the test in standardized certification programs.

Acknowledgments

This work was supported by the Centre for Clinical Education of the University of Copenhagen and the Capital Region of Denmark.

We thank Torben R. Rasmussen, MD; Birgitte H. Folkersen, MD; Peder Fabricius, MD; Lemke Pronk, MD; Anthonie van der Wekken, MD; Hans in't Veen, MD; Maren Schuhmann, MD; Dr. Claussen, MD; Aleš Rozman, MD, and Therese Lapperre, MD, for their contribution to validity evidence gathering.

Financial Disclosure and Conflicts of Interest

The authors declare that they have no competing interests.

References

- Silvestri GA, Gonzalez AV, Jantz MA, Margolis ML, Gould MK, Tanoue LT, et al: Methods for staging non-small cell lung cancer: diagnosis and management of lung cancer, 3rd ed – American College of Chest Physicians evidence-based clinical practice guidelines. *Chest* 2013;143(suppl 5):e211S–e250S.
- De Leyn P, Dooms C, Kuzdzal J, Lardinois D, Passlick B, Rami-Porta R, et al: Revised ESTS guidelines for preoperative mediastinal lymph node staging for non-small cell lung cancer: European Society of Thoracic Surgeons. 2013. http://www.ests.org/news/article/46/revised_ests_guidelines_on_preoperative_staging_of_lung_cancer.
- Vilmann P, Annema J, Clementsen P: Endosonography in bronchopulmonary disease. *Best Pract Res Clin Gastroenterol* 2009;23:711–728.
- Annema JT, van Meerbeeck JP, Rintoul RC, Dooms C, Deschepper E, Dekkers OM, et al: Mediastinoscopy vs. endosonography for mediastinal nodal staging of lung cancer: a randomized trial. *JAMA* 2010;304:2245–2252.
- Sharples LD, Jackson C, Wheaton E, Griffith G, Annema JT, Dooms C, et al: Clinical effectiveness and cost-effectiveness of endobronchial and endoscopic ultrasound relative to surgical staging in potentially resectable lung cancer: results from the ASTER randomised controlled trial. *Health Technol Assess* 2012;16:1–75, iii–iv.
- Vansteenkiste J, De Ruysscher D, Eberhardt WE, Lim E, Senan S, Felip E, et al: Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2013;24(suppl 6):vi89–vi98.
- Steinfort DP, Hew MJ, Irving LB: Bronchoscopic evaluation of the mediastinum using endobronchial ultrasound: a description of the first 216 cases carried out at an Australian tertiary hospital. *Intern Med J* 2011;41:815–824.
- Konge L, Annema J, Vilmann P, Clementsen P, Ringsted C: Transesophageal ultrasonography for lung cancer staging: learning curves of pulmonologists. *J Thorac Oncol* 2013;8:1402–1408.
- Fernandez-Villar A, Leiro-Fernandez V, Botana-Rial M, Represas-Represas C, Nunez-Delgado M: The endobronchial ultrasound-guided transbronchial needle biopsy learning curve for mediastinal and hilar lymph node diagnosis. *Chest* 2012;141:278–279.
- Medford AR: Learning curve for endobronchial ultrasound-guided transbronchial needle aspiration. *Chest* 2012;141:1643; author reply 1643–1644.
- Kohls-Gatzoulis JA, Regehr G, Hutchison C: Teaching cognitive skills improves learning in surgical skills courses: a blinded, prospective, randomized study. *Can J Surg* 2004;47:277–283.
- Roediger HL, Karpicke JD: Test-enhanced learning: taking memory tests improves long-term retention. *Psychol Sci* 2006;17:249–255.
- Kromann CB, Jensen ML, Ringsted C: The effect of testing on skills learning. *Med Educ* 2009;43:21–27.
- Karpicke JD, Roediger HL 3rd: The critical importance of retrieval for learning. *Science* 2008;319:966–968.
- Case SM, Swanson DB: Constructing Written Test Questions for the Basic and Clinical Sciences. Philadelphia, National Board of Medical Examiners, 2001.
- Schuwirth LW, van der Vleuten CP: ABC of learning and teaching in medicine: written assessment. *BMJ* 2003;326:643–645.
- Downing SM: Written tests; in Downing SM, Yudkowsky R (eds): *Assessment in Health Professions Education*, ed 1. New York, Routledge, 2009, pp 149–184.
- van der Vleuten CPM, Schuwirth LWT: Written assessments; in Dent JA, Harden RM (eds): *A Practical Guide for Medical Teachers*, ed 3. London, Churchill Livingstone, 2009, pp 323–331.
- Groth SS, Whitson BA, D'Cunha J, Maddaus MA, Alsharif M, Andrade RS: Endobronchial ultrasound-guided fine-needle aspiration of mediastinal lymph nodes: a single institution's early learning curve. *Ann Thorac Surg* 2008;86:1104–1109, discussion 1109–1110.
- Annema JT, Bohoslavsky R, Burgers S, Smits M, Taal B, Venmans B, et al: Implementation of endoscopic ultrasound for lung cancer staging. *Gastrointest Endosc* 2010;71:64–70.
- Haladyna TM: Guidelines for developing MC items; in Haladyna TM (ed): *Developing and Validating Multiple-Choice Test Items*, ed 3. London, Routledge, 2004.
- Strandbygaard J, Maagaard M, Larsen CR, Schouenborg L, Ottosen C, Ringsted C, et al: Development and validation of a theoretical test in basic laparoscopy. *Surg Endosc* 2013;27:1353–1359.
- Streiner DL, Norman GR: Validity; in Streiner DL, Norman GR (eds): *Health Measurement Scales: A Practical Guide to Their Development and Use*, ed 4. Oxford, Oxford University Press, 2008, pp 247–276.
- Schubert S, Ortwein H, Dumitsch A, Schwantes U, Wilhelm O, Kiessling C: A situational judgement test of professional behaviour: development and validation. *Med Teach* 2008;30:528–533.
- Graham B, Regehr G, Wright JG: Delphi as a method to establish consensus for diagnostic criteria. *J Clin Epidemiol* 2003;56:1150–1156.
- Downing SM: Statistics of testing; in Downing SM, Yudkowsky R (eds): *Assessment in Health Professions Education*, ed 1. New York, Routledge, 2009, pp 93–117.
- Cohen L, Manion L, Morrison K: Interviews; in Cohen L, Manion L, Morrison K (eds): *Research Methods in Education*, ed 5. New York, Routledge, 2000, pp 265–292.
- Keeney S, Hasson F, McKenna HP: A critical review of the Delphi technique as a research methodology for nursing. *Int J Nurs Stud* 2001;38:195–200.

- 29 Haladyna TM: Validity evidence coming from statistical study of item responses; in Haladyna TM (ed): *Developing and Validating Multiple-Choice Test Items*, ed 3. London, Routledge, 2004.
- 30 Magill RA: The stages of learning; in Magill RA (ed): *Motor Learning and Control: Concepts and Application*, ed 9. New York, McGraw-Hill, 2011.
- 31 Hawkins RE, Swanson DB: Using written examinations to assess medical knowledge and its application; in Holmboe ES, Hawkins RE (eds): *Practical Guide to the Evaluation of Clinical Competence*, ed 1. Philadelphia, Mosby, 2008, pp 42–59.
- 32 Streiner DL, Norman GR: Dividing the items; in Streiner DL, Norman GR (eds): *Health Measurement Scales: a Practical Guide to Their Development and Use*, ed 4. Oxford, Oxford University Press, 2008.
- 33 Miller GE: The assessment of clinical skills/competence/performance. *Acad Med* 1990; 65(suppl 9):S63–S67.
- 34 Nunnink L, Venkatesh B, Krishnan A, Vidhani K, Udy A: A prospective comparison between written examination and either simulation-based or oral viva examination of intensive care trainees' procedural skills. *Anaesth Intensive Care* 2010;38:876–882.
- 35 Konge L, Vilmann P, Clementsen P, Annema JT, Ringsted C: Reliable and valid assessment of competence in endoscopic ultrasonography and fine-needle aspiration for mediastinal staging of non-small cell lung cancer. *Endoscopy* 2012;44:928–933.
- 36 Davoudi M, Colt HG, Osann KE, Lamb CR, Mullon JJ: Endobronchial ultrasound skills and tasks assessment tool: assessing the validity evidence for a test of endobronchial ultrasound-guided transbronchial needle aspiration operator skill. *Am J Respir Crit Care Med* 2012;186:773–779.
- 37 Kemp SV, El Batrawy SH, Harrison RN, Skwarski K, Munavvar M, Rosell A, et al: Learning curves for endobronchial ultrasound using CUSUM analysis. *Thorax* 2010; 65:534–538.
- 38 Konge L, Annema J, Clementsen P, Minddal V, Vilmann P, Ringsted C: Using virtual-reality simulation to assess performance in endobronchial ultrasound. *Respiration* 2013;86: 59–65.