Respiration

Respiration , DOI: 10.1159/000546873
Received: March 21, 2025
Accepted: June 4, 2025
Published online: June 12, 2025

Development and validation of two bronchoscopy knowledge assessments

Gerretsen ECF, Popeijus HE, Annema JT, Clementsen PF, Corbetta L, Gompelmann D, van der Heijden EHFM, Hiddinga BI, van Mook WNKA, Munavvar M, Smeenk FWJM, Groenier M

ISSN: 0025-7931 (Print), eISSN: 1423-0356 (Online)
https://www.karger.com/RES
Respiration

Research Article

Development and validation of two bronchoscopy knowledge assessments

Eveline C F Gerretsen[a], Herman E Popeijus[b], Jouke T Annema[c], Paul F Clementsen[d], Lorenzo Corbetta[e], Daniela Gompelmann[f], Erik H F M van der Heijden[g], Birgitta I Hiddinga[h], Walther N K A van Mook[i,j,k], Mohammed Munavvar[l,m], Frank W J M Smeenk[k,n], Marleen Groenier[o]

[a] Department of Educational Development and Research, School of Health Professions Education (SHE), Maastricht University, Maastricht, the Netherlands
[b] Institute of Nutrition and Translational Research in Metabolism, Department of Nutrition and Movement Sciences, Faculty of Health, Medicine and Life Sciences, Maastricht University, the Netherlands
[c] Department of Respiratory Medicine, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, the Netherlands
[d] Copenhagen Academy for Medical Education and Simulation, Copenhagen, Denmark
[e] Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy
[f] Division of Pulmonology, Department of Internal Medicine II, Medical University of Vienna, Vienna, Austria
[g] Department of Respiratory Medicine, Radboudumc, Nijmegen, the Netherlands
[h] Department of Pulmonary Medicine and Tuberculosis, University Medical Center Groningen, Groningen, the Netherlands
[i] Department of Intensive Care, Maastricht University Medical Center+, Maastricht, the Netherlands
[j] Academy for Postgraduate Training, Maastricht University Medical Center+, Maastricht, the Netherlands
[k] School of Health Professions Education (SHE), Maastricht University, Maastricht, the Netherlands
[l] Respiratory Medicine, Lancashire Teaching Hospitals, Preston, United Kingdom
[m] University of Central Lancashire, Preston, United Kingdom
[n] Department of Medical Education, Catharina Hospital, Eindhoven, the Netherlands
[o] Department of Technical Medicine, University of Twente, Enschede, the Netherlands

Short Title: Bronchoscopy knowledge assessment: development and validation

Corresponding Author:
Eveline Gerretsen
e.gerretsen@maastrichtuniversity.nl

Keywords: bronchoscopy; knowledge; assessment; development; validation

Word count: 3749 words

Abstract
Introduction: Simulation-based training (SBT) is a key method for teaching bronchoscopy skills to pulmonology residents. A theoretical foundation can enhance SBT efficiency. This study developed and evaluated the validity of an anatomy and theoretical bronchoscopy exam using Kane's validity framework.
Methods: 19 anatomy and 58 theoretical exam questions, developed by pulmonology experts, were assessed through two Delphi rounds. Both exams were then taken by 53 prepared pulmonology residents. The theoretical exam was also taken by three unprepared groups: novices, intermediates and experts. Using the residents' data, scoring evidence for the theoretical exam was evaluated using item difficulty and item discrimination indices, and generalization evidence was assessed using Cronbach's alpha. Extrapolation evidence was obtained by comparing theoretical exam scores across the different groups. Implications evidence for both exams was gathered by evaluating residents' preparedness, based on exam performance and instructor feedback.
Results: The Delphi procedure resulted in 19 anatomy and 31 theoretical questions. Item difficulty values predominantly ranged from 0.85-1.0, item discrimination indices mostly ranged from 0.0-0.25. Cronbach's alpha was 0.55. While scores appeared to correlate with experience, no significant differences were observed between the four groups. Most residents passed both exams on their first attempt, and instructors rated their anatomical knowledge as good.
Conclusion: Expert involvement and acceptable item difficulty, item discrimination and internal consistency supported the exams' validity. The exams also effectively motivated residents to prepare for SBT. These findings highlight the value of pre-SBT exams in enhancing residents' preparation, allowing more time to focus on mastering procedural skills.

Introduction

Over the past two decades, simulation-based training has increasingly been employed to teach flexible bronchoscopy skills to pulmonology trainees [1]. This approach allows trainees to practice procedural skills without compromising patient safety, making it a desirable training alternative to the traditional apprenticeship model. However, while simulation-based training is effective for developing practical skills, it does not provide trainees with a comprehensive understanding of knowledge such as topical anesthesia, sedation, anatomy, complications and (contra-)indications for the procedure. Recognizing this limitation, pulmonology educators have emphasized that a theoretical stage in bronchoscopy training should precede simulation-based training [2]. Pre-existing knowledge in general, and especially anatomical knowledge of the bronchial tree, can potentially enhance the efficiency of simulation-based training by reducing the time spent on theoretical instruction and explanation of anatomy, allowing trainees to focus more effectively on mastering procedural skills.

Basic knowledge relevant for bronchoscopy training is typically assessed through written examinations. Despite the recognized importance of these assessments, few validation studies have been conducted on theoretical bronchoscopy exams [3-5]. Moreover, these studies have significant limitations: one, conducted 16 years ago [3], may no longer fully reflect current clinical practices, while the others provided limited validity evidence, relying solely on expert consensus for item development [4], and another conducted item analysis on a sample of only seven participants [5]. These limitations highlight the need for a more rigorous validation process of knowledge assessments. In assessment, validity is not simply about whether a test accurately measures a specific construct, but rather about the strength of the evidence supporting the interpretations and uses of exam scores [6]. Comprehensive validity assessment requires triangulating evidence from several sources [7]. In this study, we will use Kane's validity framework, which emphasizes that validity is a chain of inferences, each requiring evidence to be collected [6]. These inferences include *scoring*, referring to the appropriateness of scoring criteria, *generalization*, concerning the generalizability of the test scores to the broader domain, *extrapolation*, concerning the meaning of the test score for real-life performance and *implications*, concerning the use of the test scores to make decisions about learners [8].

In summary, there is a scarcity of validity evidence for exams assessing the knowledge required for bronchoscopy performance. This study examined the validity of an anatomy and a theoretical exam in bronchoscopy, gathering evidence for the *scoring*, *generalization, extrapolation* and *implications* inferences. As such, this study intended to provide valuable insights for future development and refinement of knowledge assessments for pulmonology residents.

Methods

Context

In 2020, a mandatory simulation-based training program for novice Dutch pulmonology residents was introduced [9]. To ensure that residents were adequately prepared, two exams were implemented as prerequisites for participation in the training program. These exams were developed by a panel of six Dutch pulmonology experts. The first exam, hereafter referred to as the 'anatomy exam' for clarity, focused exclusively on anatomy. A 100% score was required to ensure that residents entering the simulation-based bronchoscopy training had adequate anatomical knowledge of the bronchial tree, so that no valuable training time would be wasted on explaining anatomy. The second exam, hereafter referred to as the 'theoretical exam', was based on the British Thoracic Society (BTS) guideline [10]. This exam evaluated broader bronchoscopy-related knowledge, including topical anesthesia, sedation, pre-procedure preparations, (contra-)indications for bronchoscopy, monitoring, complications, staff and hygiene protocols, and sampling techniques. Trainees were required to answer 60% of the questions correctly on the theoretical exam. Since the most important requirement was that residents had a solid understanding of anatomical knowledge

before entering the simulation-based training program, broader bronchoscopy-related knowledge was considered less critical for admission. These entry requirements aimed to ensure that residents prepared thoroughly, entering the simulation-based training with a solid knowledge base.

Development of the exams

The initial anatomy and theoretical exam, developed by the Dutch pulmonology experts, contained 19 and 58 multiple choice questions, respectively. These questions were then reviewed by a test expert (HP), who identified issues related to clarity, phrasing and question structure. Questions were refined to ensure they were unambiguous and well-structured. Next, two Delphi rounds were conducted with five European pulmonology experts (PC, LC, DG, BH, MM). In each round, the experts rated the relevance of each question using a 5-point Likert scale and provided feedback where necessary. A question was considered relevant if at least four experts rated it as 4 or 5 out of 5. During the first Delphi round, all 19 anatomy questions and 20 theoretical questions were consistently rated as relevant by the experts and were directly included in the final question sets. The remaining 38 theoretical questions were judged by the researchers, who determined that three questions were ambiguous and had to be discarded. The remaining 35 theoretical questions were re-evaluated in a second Delphi round. Following this second review, 11 additional theoretical questions were deemed relevant. Ultimately, this process resulted in the inclusion of 19 questions in the anatomy exam and 31 questions in the theoretical exam. One example question from each exam is provided in Online Supplementary Material 1 to illustrate the structure and content of the items.

Participants and data collection

The anatomy and theoretical exams were taken by Dutch pulmonology residents (n = 53), hereafter referred to as 'residents' for clarity, who were required to pass both before being allowed to attend the simulation-based training. The residents were instructed to study material related to bronchoscopy [10-13].

The theoretical exam was also administered to three additional groups of participants who were explicitly instructed not to prepare for the test: 1) medical residents without any bronchoscopy experience (novices; n = 15), 2) pulmonology residents who started their residency before 2020 and, therefore, did not have to attend the mandatory simulation-based training program and had performed 5–100 bronchoscopies (intermediates; n = 13) and 3) pulmonologists who had performed more than 500 bronchoscopies (experts; n = 14). Only the theoretical exam was administered to these groups, as the anatomy exam was deemed irrelevant: novices, having no anatomical knowledge, would achieve scores close to 0%, while those regularly performing bronchoscopies would be expected to score near 100%. Consequently, no additional *scoring*, *generalization*, or *extrapolation* evidence was collected for the anatomy exam, as no meaningful differences were expected. Instead, only *implications* evidence was collected for the anatomy exam. Table 1 shows demographics of all participant groups.

Participants provided written informed consent before receiving a link to the online exam environment via email, which allowed them to complete the exams at their convenience. Due to logistical constraints, no formal testing conditions, supervision, or remote proctoring were implemented. Both exams were administered through the online testing platform *Remindo* (version 22.5 to 24.4) and responses were collected between October 2022 and October 2024. The anatomy exam consisted of 19 questions, from which 10 were randomly selected for each resident. A passing score of 100% was required. In case of a retake, a new set of 10 questions was randomly selected from the original 19. The theoretical exam, which covered all 31 questions, required a passing score of 60%. Residents retaking the theoretical exam were presented with the same 31 questions as in their initial attempt. To minimize the possibility of extensive searching by the participants, the time to complete the exams was limited to one hour. For all other participant groups, only data from those completing both the anatomy and theoretical exams within this one-hour timeframe on their

first attempt were included in the analysis. The dataset contained no personally identifiable information; researchers could only access participants' experience levels, responses to the questions and their final scores. This study was approved by the Ethics Committee of the University of Twente (approval number 210232).

Scoring evidence

To evaluate item quality of the theoretical exam, item difficulty and item discrimination were calculated using the data of the residents who prepared for the exam, considering only their first attempt. Item difficulty was determined by calculating the percentage of residents who answered each item correctly; a lower value may indicate a higher difficulty for that item. Item discrimination was analyzed by ranking residents based on their final scores. Following the methodology of similar studies [14,15], the resident data were divided into three groups: the 17 residents with the lowest scores, the 19 residents with the middle scores and the 17 residents with the highest scores. The discrimination index for each item was calculated as the difference in the proportion of correct answers between the high- and low-scoring groups, indicating how well each question differentiated between stronger and weaker performers. A higher discrimination index indicates that the item more effectively distinguishes between high- and low-performing individuals. All analyses were conducted in RStudio, R version 4.4.1.

Generalization evidence

Internal consistency of the theoretical exam, using the data of the residents who prepared for the exam and considering only their first attempt, was assessed primarily using Cronbach's alpha, calculated with the Pysch package in R. Values closer to 1.0 indicate strong internal consistency, meaning that exam items are highly correlated and measure a similar underlying concept. However, in a highly homogeneous group with little variance in exam scores, internal consistency may decrease because the items lose their ability to differentiate between examinees, reducing the overall covariance between items. Therefore, alternative reliability metrics were also calculated as supplementary analyses: split-half reliability, using the Spearman-Brown formula, and McDonald's omega total.

Extrapolation evidence

Theoretical exam scores, considering only participants' first attempts, were compared across the four participant groups using the Kruskal-Wallis test in RStudio. Post-hoc pairwise comparisons were performed with Bonferroni adjustments to account for multiple comparisons. The effect size for the overall group difference was estimated using Epsilon squared (E^2) [16]. Additionally, a Spearman's rank-order correlation was performed to examine the relationship between bronchoscopy experience level (novice, intermediate, expert) and theoretical exam scores.

Implications evidence

The most important objective was to ensure that, before starting the simulation-based bronchoscopy training, residents would have acquired adequate anatomical knowledge of the bronchial tree. The training program began with a brief 15-minute recap of bronchial anatomy by the trainer, intended as a refresher rather than instruction. As residents were required to achieve a perfect score on the exam prior to participation, we recorded how many attempts were needed to pass the anatomy exam, as this may provide insight into how thoroughly residents prepared for the exam. We also administered a questionnaire to the instructors of the simulation-based training, asking them to 1) rate how well-prepared residents were on average regarding anatomical knowledge, on a Likert scale from 1 to 5, with higher ratings indicating greater preparedness, and 2) indicate how much additional time, on average, was needed to explain anatomy during the training, beyond the initial recap. This was rated on a scale where 1 = 0 minutes, 2 = 15 minutes, 3 = 30 minutes, 4 = 1 hour and 5 = 2 additional hours. Furthermore, to gather evidence for the *implications* inference of the theoretical

exam, we compared the proportion of residents who prepared for the exam and passed on their first attempt with the proportions of the participants in the other unprepared groups.

Results

Scoring evidence

Based on the data of the residents who prepared for the exam (n = 53), item difficulty values for the theoretical exam varied from 0.3 to 1.0, with a notable peak in the 0.85-1.0 range, shown in Fig. 1). The item discrimination indices ranged from -0.23 to 0.59, with the distribution peaking in the 0.0 – 0.25 range (shown in Fig. 2).

Generalization evidence

The internal consistency of the theoretical exam, assessed using Cronbach's alpha of the data of the residents who prepared for the exam, was 0.55. Supplementary analyses showed a Spearman-Brown corrected split-half reliability of 0.61 and McDonald's omega total of 0.76.

Extrapolation evidence

A comparison of performance on the theoretical exam across the four participant groups revealed a significant difference (Kruskal-Wallis H(3) = 51.95, $p < 0.001$, $\varepsilon^2 = 0.54$). Post hoc pairwise comparisons with Bonferroni correction showed that residents scored significantly higher than novices ($Z = -6.04$, $p < 0.001$), intermediates ($Z = -4.31$, $p < 0.001$) and experts ($Z = -4.11$, $p < 0.001$). Median scores and interquartile ranges for each group (maximum score = 31) are shown in Figure 3. Although experts scored slightly higher than intermediates ($Z = -0.26$, adjusted p = 1.000) and intermediates scored higher than novices ($Z = -1.14$, adjusted p = 1.000), these differences were not statistically significant. The same was true for the comparison between novices and experts ($Z = -1.42$, adjusted p = 0.93). However, a Spearman's rank-order correlation revealed a statistically significant positive association between the level of bronchoscopy experience and theoretical exam scores ($\rho = 0.49$, $p = 0.001$).

Implications evidence

37 residents (69.8%) passed the anatomy exam at their first attempt, 13 (24.5%) passed the exam at their second attempt and three (5.7%) passed the exam at their third attempt. Regarding the questionnaire, ten instructors (83.3%) responded. Nine instructors rated the residents' preparedness for the simulation-based training in terms of anatomical knowledge as 4 out of 5 and one instructor rated their preparedness as 5 out of 5. As one instructor spontaneously added: "I found most candidates to be very well-prepared and highly motivated to follow the training program". Most instructors (n = 5) required on average per training session an additional 15 minutes to explain anatomy, while two needed no additional time, one required 30 minutes, and two required 1 hour. Table 2 shows failure rates for the theoretical exam across the different groups. Residents had a lower failure rate on the theoretical exam (1.8%, 95% CI: 0.3-9.9%) compared to the unprepared participants (64.3%, 95% CI: 49.2-77.0%). When broken down by subgroup, failure rates were highest among novices (86.7%, 95% CI: 62.1-96.3%), followed by intermediates (69.2%, 95% CI: 42.4-87.3%) and experts (35.7%, 95% CI: 16.3-61.2%).

Discussion

This study explored the validity of an anatomy and a theoretical bronchoscopy exam. Kane's validity framework was used as a guiding structure for evaluating the validity of the exams, and multiple sources of evidence were gathered to assess their validity. The *scoring* inference for both exams was supported by the expert-driven development of questions and their refinement by a test expert. For the theoretical exam, this inference was further supported by the observed item difficulty and item discrimination indices. Evidence relevant to the *generalization* inference of the theoretical exam was

provided by a Cronbach's alpha of 0.55. The e*xtrapolation* inference for both exams was supported by the Delphi process, in which a panel of international experts assessed question relevance. Additionally, evidence for this inference for the theoretical exam was provided by comparing test performance across different participant groups. Finally, regarding the *implications* inference, the results suggested that both exams effectively contributed to residents' preparedness for simulation-based training. This was evidenced by the high pass rates, with the vast majority of residents passing the anatomy exam on their first attempt and almost all passing the theoretical exam on their first attempt, in contrast to the higher failure rates for the theoretical exam observed in the other unprepared groups. Additionally, instructors rated the residents' anatomical knowledge as generally good at the start of the training, with most requiring only a small amount of time during the simulation-based training to explain anatomy.

A large proportion of items in the theoretical exam had low discrimination indices, with 19 questions scoring below 0.2. While previous studies on the development of theoretical tests in the field of surgical endoscopy and endosonography [17-18] consistently excluded such questions, we do not regard low item discrimination indices as problematic in our study. Given that nearly all residents passed, it is expected that the questions would not differentiate well between high- and low-performing individuals. This exam was designed to ensure that residents would have adequate (especially anatomy) knowledge before beginning simulation-based bronchoscopy training, rather than to differentiate between varying performance levels on a theoretical exam. This is also reflected in the higher item difficulties in our study (i.e., questions were relatively easy, with the majority of questions having item difficulties between 0.85 and 1.0), whereas in the other studies [17-18], most questions fell within the middle difficulty range (i.e., item difficulty between 0.45 and 0.75). The observed difficulty levels reflect both the nature of the examinee population (a highly selected and motivated group of residents specializing further) and the exam's intent to assess essential knowledge rather than differentiate performance levels. Given these factors, the combination of expert-driven question development, item difficulty, and discrimination values provides strong support for the validity of the scoring inference.

One important aspect of exam quality is whether the items collectively measure the intended construct, which in this study was assessed using Cronbach's alpha. The Cronbach's alpha value for the theoretical exam was 0.55, somewhat below the commonly accepted threshold of 0.7 [19] and lower than the values reported in the other two above-mentioned studies (i.e., 0.75 and 0.91). While a Cronbach's alpha of 0.55 is typically considered 'poor', this was expected given the small number of questions in the exam [20,21], making it an acceptable value of internal consistency for such a short exam. Increasing the number of questions might improve internal consistency and strengthen the *generalization* inference. However, expanding the question set was not feasible due to the limited scope of the subject matter and the risk of excessive repetition. Notably, when combining data from all participant groups, Cronbach's alpha increased to 0.86, indicating strong internal consistency when applied to a more heterogeneous population. When examining groups separately, Cronbach's alpha was 0.80 for the novices, 0.72 for the intermediates and 0.68 for the experts. These findings align with expectations, as participants in these groups had not studied the specific exam material, resulting in greater score variance and consequently, higher internal consistency. Additionally, McDonald's omega total (0.76) and the Spearman-Brown corrected split-half reliability (0.61) suggest a moderate internal consistency, supporting the interpretation that the exam items measured the intended construct.

Although the scores for the theoretical exam slightly improved from novices to intermediates and experts, no significant differences in test performance were observed between these groups. This lack of significance is most likely due to the small group sizes, rather than the absence of differences. Additionally, somewhat surprisingly, 36% of experts failed the exam. However, this outcome is unlikely due to test content irrelevance, as the exam was developed by experts and questions

underwent relevance assessment by the international panel. Instead, this finding might be due to different practices across centers, where experts might be adhering to local protocols rather than the BTS guideline. This discrepancy highlights the limitation of relying on experts making a test as a source of validity evidence in these circumstances. Nevertheless, the observed performance differences between the participant groups, despite the small sample sizes, provide some support for the *extrapolation* inference, as the findings suggest that test scores reflect underlying differences in bronchoscopy-related knowledge.

The *implications* inference is considered the most critical [6], as it evaluates whether the exams achieved their intended purpose of ensuring that residents where adequately prepared for simulation-based training. The high pass rates for the theoretical and anatomy exam suggest that the requirement to pass the exams motivated residents to prepare. Additionally, instructor feedback indicated that residents generally had sufficient anatomical knowledge at the start of the training, and little training time was spent on anatomy instruction. However, since there was no control group of residents who did not make the exam but still participated in the training program, it remains uncertain to what extent the anatomy exam itself contributed to the residents' preparedness. Therefore, while these findings provide some support for the *implications* inference, the strength of the evidence remains somewhat limited.

Strengths and limitations

We believe this study has several strengths. First, the exams were thoroughly developed by Dutch pulmonology experts and refined through input from five international experts and one test expert. This ensured a high level of consensus regarding question relevance and clarity. Item discrimination and item difficulty values further supported the strength of the *scoring* inference. Second, the theoretical exam successfully differentiated between participants who prepared for the exam and those who did not, demonstrating its sensitivity to the knowledge acquisition through preparation materials. Third, while no significant differences were observed between novices, intermediates, and experts, a separate Spearman's correlation analysis demonstrated a significant positive relationship between level of expertise and theoretical exam scores (Spearman's $\rho = 0.49$, $p = 0.001$), providing some support for the *generalization* inference. Finally, the high pass rate among pulmonology residents demonstrates that the exam achieved its primary purpose: ensuring participants met the minimum knowledge requirements necessary for participation in the practical simulation-based training.

Despite these strengths, this study also has limitations. First, the theoretical exam included a limited number of questions, which may have limited its internal consistency, consequently limiting the strength of the *generalization* inference. Second, the evidence for the *implications* inference for the anatomy exam in this study was also somewhat weak due to the lack of a control group. Future studies could benefit from incorporating designs that allow for stronger inferences regarding the impact of examination of anatomical knowledge on residents' preparedness during simulation-based training. Third, while ideal testing conditions would have included proctoring, logistical constraints precluded remote supervision of the online exams. Consequently, this absence of controlled testing conditions might have led to an overestimation of anatomical and theoretical knowledge if participants actually used external resources, such as the internet, during the exam. Finally, the strict adherence to the BTS guideline in designing the exam may have reduced its alignment with Dutch clinical practice, potentially impacting the performance of experts who may be familiar with other treatment protocols used in their own hospitals. To improve alignment with actual clinical practice, locally used guidelines should be taken into account when developing future versions of the exam, rather than relying solely on international standards.

Conclusion

This study evaluated the effectiveness of an anatomy and a theoretical exam in ensuring pulmonology residents' readiness for simulation-based training and evaluated their validity using Kane's framework. The validity of the exams was supported by expert involvement in both their design and refinement, combined with acceptable item difficulty, item discrimination indices and internal consistency. The findings suggest that the exams effectively motivated residents to prepare thoroughly, contributing to their acquisition of adequate anatomical knowledge prior to entering the simulation-based bronchoscopy training. These results highlight the value of using exams as entry requirements for simulation-based training, as they enhance residents' preparation, potentially allowing for more time to focus on mastering procedural skills.

Statements

Acknowledgement
The authors would like to thank the Dutch Bronchoscopy Simulation (DBS) Study Group (AF Aldenkamp, E Citgez, LMMJ Crombag, W Hagmolen of ten Have, BPC Hoppe, MK Ninaber, MA van de Pol, B Robberts, M Rutten, R Sprooten, M Wagenaar).

Statement of Ethics
Study approval statement: This study protocol was reviewed and approved by the Ethics Committee of the University of Twente, approval number 210232.
Consent to participate statement: Participants provided written informed consent before receiving a link to the online exam environment via email.

Conflict of Interest Statement
DG obtained lecture fees from Olympus, Pulmonx, AstraZeneca, Boehringer Ingelheim, MSD, Chiesi and BerlinChemie. Erik H. van der Heijden and Mohammed Munavvar were both members of the journal's Editorial Board at the time of submission. EH has received unrestricted research grants paid to his institution from Astra Zeneca Oncology, Pentax Medical, Philips Image Guided Therapies, Johnson & Johnson, Intuitive, IHI-EU public fund (for the "IMAGIO" project), and KWF–National Cancer Fund (for a KWF implementation research project). He has also received consulting fees paid to his institution from Johnson & Johnson, Intuitive, and NLC. In addition, EH has received speaker's fees paid to his institution from Janssen-Cilag and Siemens, and sponsorship for symposiums he organized from Pentax, Ethicon, Astra Zeneca, Intuitive, and Philips Image Guided Therapy. Travel fees have been paid to EH's institution by Siemens. He holds issued, pending, and planned patents in the field of advanced and navigation bronchoscopy. He has received single-use instruments for a scientific research protocol from Philips Medical and equipment loans for scientific research from Intuitive Surgical and Pentax Medical. EH is also an unpaid member of the Executive Board of EABIP. The other authors have no conflicts of interest to declare.

Funding Sources
The PhD project of Eveline Gerretsen was funded by the Catharina Hospital research fund and the board of directors of Maastricht University Medical Center+. The funder had no role in the design, data collection, data analysis, and reporting of this study.

Author Contributions
PC, LC, DG, BH and MM participated in the Delphi procedure. EG, JA, EH, WM, FS, MG contributed to the study design. EG conducted the data analysis and interpretation. HP, JA, EH, WM, FS and MG contributed to data interpretation. All authors contributed to writing and/or critically reviewing the manuscript.

Data Availability Statement
Data associated with this study is deposited in DataverseNL, https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/KEW82V. The exam questions are not publicly available, as pulmonology residents will need to take the exams before the simulation-based training. However, the questions can be provided by the corresponding author upon request.

References [Numerical]

1. Gerretsen EC, Chen A, Annema JT, Groenier M, van der Heijden EH, van Mook WN, et al. Effectiveness of Flexible Bronchoscopy Simulation-Based Training: A Systematic Review. Chest. 2023;164(4):952-62.
2. Corbetta L, Arru LB, Mereu C, Pasini V, Patelli M. Competence and training in interventional pulmonology. Panminerva Med. 2019;61(3):203–31.
3. Quadrelli S, Davoudi M, Galíndez F, Colt HG. Reliability of a 25-Item Low-Stakes Multiple-Choice Assessment of Bronchoscopic Knowledge. Chest. 2009;135(2):315–21.
4. Davoudi M, Quadrelli S, Osann K, Colt HG. A competency-based test of bronchoscopic knowledge using the Essential Bronchoscopist: an initial concept study. Respirology. 2008;13(5):736-43.
5. Colt HG, Davoudi M, Murgu S, Rohani NZ. Measuring learning gain during a one-day introductory bronchoscopy course. Surg Endosc. 2011;25(1):207-16.
6. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ. 2015;49(6):560–75.
7. Sullivan GM. A Primer on the Validity of Assessment Instruments. J Grad Med Educ. 2011;3(2):119–20.
8. Carrillo-Avalos BA, Leenen I, Trejo-Mejía J, Sánchez-Mendiola M. Bridging Validity Frameworks in Assessment: Beyond Traditional Approaches in Health Professions Education. Teach Learn Med. 2023;1-10.
9. Gerretsen EC, Groenier M, Annema J, Heijden EH, van Mook WN, Aldenkamp AF, et al. Basic Bronchoscopy Competence Achieved by a Nationwide One-day Simulation-based Training. J Bronchology Interv Pulmonol. 2025;32(1):e0995.
10. Du Rand IA, Blaikley J, Booton R, Chaudhuri N, Gupta V, Khalid S, et al. British Thoracic Society guideline for diagnostic flexible bronchoscopy in adults. Thorax. 2013;68:i1-i44.
11. Bonta PI, Koster D, Slebos DJ. NVALT. LEIDRAAD Bronchoscopie. [Internet]. Available from: https://www.nvalt.nl/aios/bronchoscopie-cursus/_/Studiemateriaal/LEIDRAAD%20Bronchoscopie%20versie%2026-11-2018.pdf. [Accessed November 26th 2024].
12. Lung Segments. Nuclear Medicine Information. [Internet]. Available from: http://www.nucmedinfo.com/Pages/lungsegmentsbase.html. [November 26th 2024].
13. Clementsen PF, Nayahangan LJ, Konge L. NVALT. Bronchoscopy A Practical Handbook. [Internet]. Available from: https://www.nvalt.nl/aios/bronchoscopie-cursus/_/Studiemateriaal/handbook%20bronchoscopy%20CAMES_EvdH.pdf. [Accessed November 26th 2024].
14. Bhat SK, Prasad KH. Item analysis and optimizing multiple-choice questions for a viable question bank in ophthalmology: A cross-sectional study. Indian J Ophthalmol. 2021;69(2):343–6.
15. Cappelleri JC, Jason Lundy J, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. Clin Ther. 2014;36(5):648–62.
16. In J, Lee Dk. Alternatives to the P value: connotations of significance. Korean J Anesthesiol. 2024;77(3):316-25.
17. Savran MM, Hansen HJ, Petersen RH, Walker W, Schmid T, Bojsen SR, et al. Development and validation of a theoretical test of proficiency for video-assisted thoracoscopic surgery (VATS) lobectomy. Surg Endosc. 2015;29(9):2598–604.
18. Savran MM, Clementsen PF, Annema JT, Minddal V, Larsen KR, Park YS, et al. Development and validation of a theoretical test in endosonography for pulmonary diseases. Respiration. 2014;88(1):67–73.
19. Taber KS. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. Res Sci Educ. 2018;48:1273–96.
20. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. 2011;2:53–5.

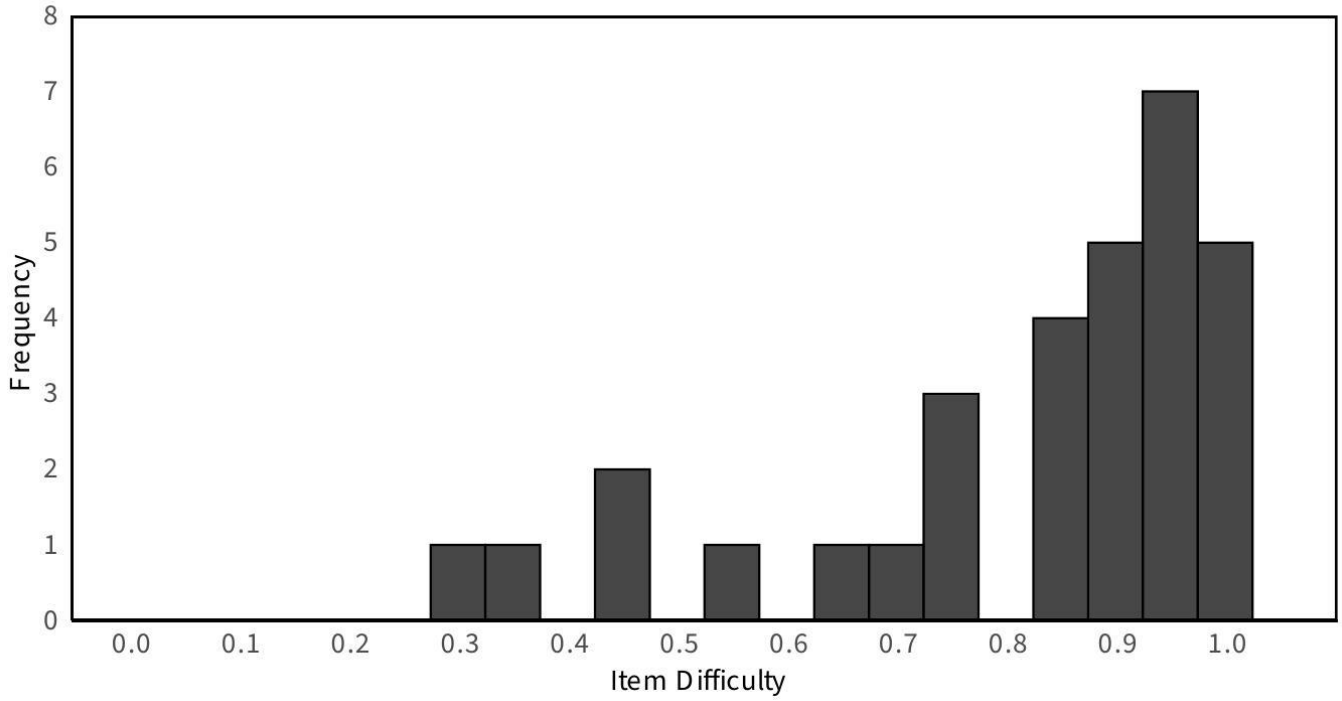
21. Trinchera L, Marie N, Marcoulides GA. A Distribution Free Interval Estimate for Coefficient Alpha. Struct Equ Model. 2018;25(6):876–87.

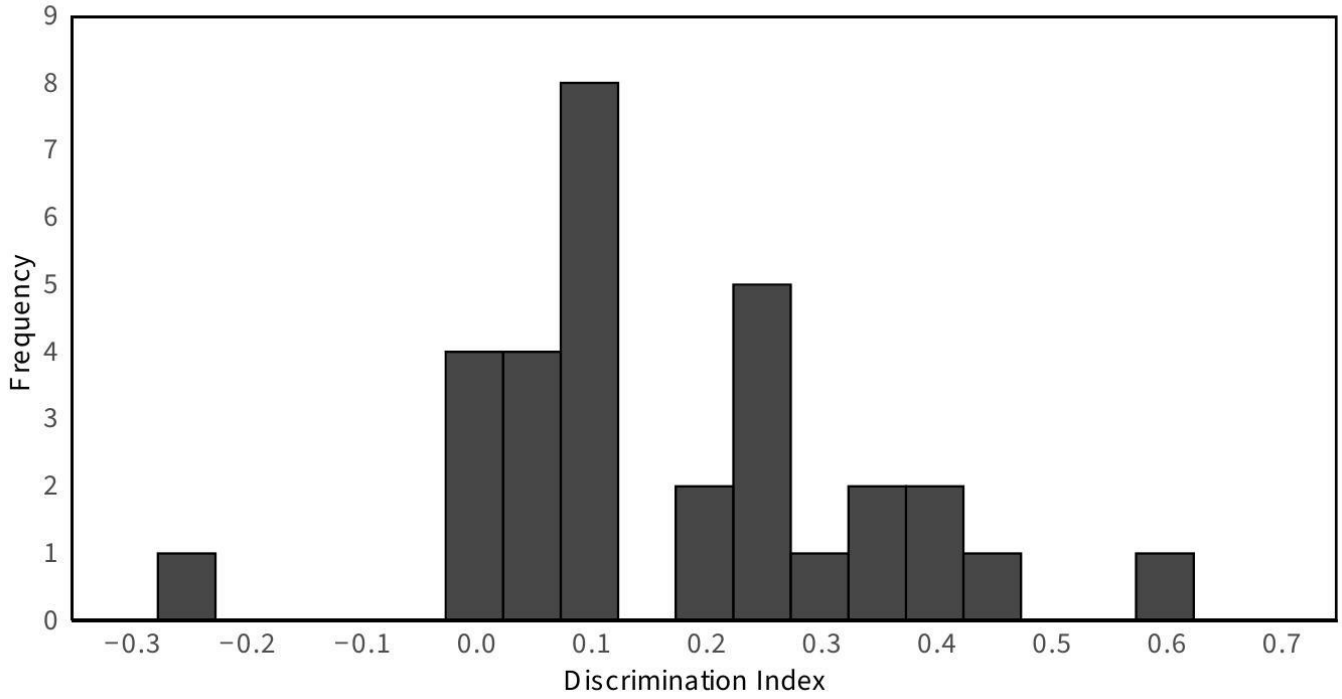
Figure Legends

Fig. 1. Distribution of item difficulty indices for the theoretical exam (a higher score indicating an 'easier' question), based on the exam results of the residents (n = 53). The Y-axis represents the number of questions.

Fig. 2. Distribution of item discrimination indices for the theoretical exam (with a higher index indicating a better discrimination ability between 'high' and 'low' performing participants), based on the exam results of the residents (n = 53). The Y-axis represents the number of questions.

Fig. 3. Comparison of final scores for the theoretical exam between novices, intermediates, experts and residents (max = 31).





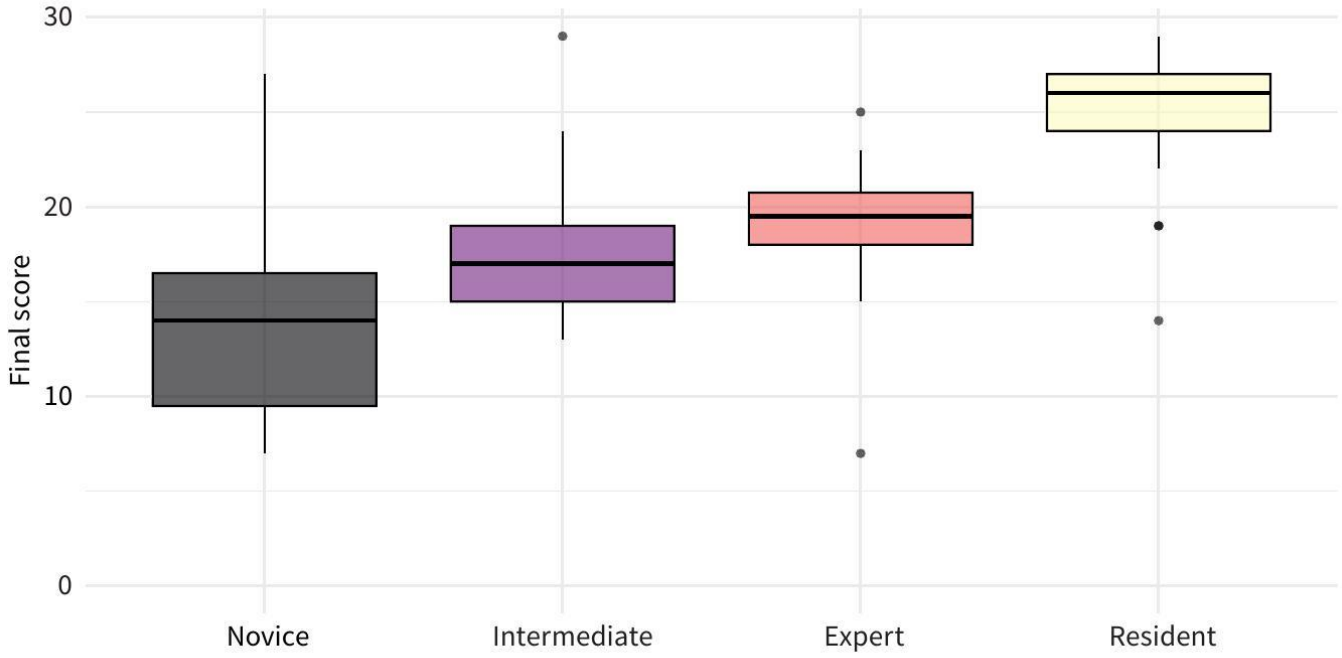


Table 1. Participant demographics

Group	N	Age (mean±SD)	N male (%)	N female (%)
Novices*	15	28.0 (±3.1)	6 (40)	9 (60)
Intermediates**	13	31.2 (±3.4)	6 (46)	7 (54)
Experts***	14	42.3 (±8.2)	8 (57)	6 (43)
Residents******	53	30.9 (±2.9)	18 (34)	35 (66)

* = non-pulmonology residents without any bronchoscopy experience, ** = pulmonology residents who started their residency before 2020, had performed 5–100 bronchoscopies, and did not have to attend the mandatory simulation-based training program, *** = pulmonologists who had performed more than 500 bronchoscopies, **** = pulmonology residents who had to attend the mandatory simulation-based training program and were required to pass both the anatomy and theoretical exam.

Table 2. Theoretical exam failure rates per group

Group	Failed (N, %)
Novice	13 (86.7%)
Intermediate	9 (69.2%)
Expert	5 (35.7%)
Resident	1(1.9%)