Thieme

Validity evidence for endoscopic ultrasound competency assessment tools: Systematic review

OPEN ACCESS

Authors
Alessandra Ceccacci[1], Harneet Hothi[2], Rishad Khan[3], Nikko Gimpaya[4], Brian P.H. Chan[5,1,3], Nauzer Forbes[6], Paul James[7,1,3], Daniel Jeffry Low[5,1,3], Jeffrey Mosko[8,1,3,9], Elaine T. Yeung[5,1,3], Catharine M Walsh[‡10,11], Samir C Grover[‡1,5,3,4]

Institutions
1 Department of Medicine, University of Toronto, Toronto, Canada
2 Temerty Faculty of Medicine, University of Toronto, Toronto, Canada
3 Division of Gastroenterology and Hepatology, University of Toronto, Toronto, Canada
4 Scarborough Health Network Research Institute, Scarborough Health Network, Scarborough, Canada
5 Division of Gastroenterology, Scarborough Health Network, Scarborough, Canada
6 Division of Gastroenterology and Hepatology, University of Calgary, Calgary, Canada
7 Division of Gastroenterology, University Health Network, Toronto, Canada
8 Division of Gastroenterology, St Michael's Hospital, Toronto, Canada
9 Li Ka Shing Knowledge Institute, Unity Health Toronto, Toronto, Canada
10 Division of Gastroenterology, Hepatology, and Nutrition, and the Research and Learning Institutes, The Hospital for Sick Children, Toronto, Canada
11 Department of Pediatrics and the Wilson Centre, University of Toronto Temerty Faculty of Medicine, Toronto, Canada

Key words
Endoscopic ultrasonography, Quality and logistical aspects, Training, Performance and complications

received 19.10.2024
accepted after revision 5.11.2024
accepted manuscript online 11.11.2024

Bibliography
Endosc Int Open 2024; 12: E1465–E1475
DOI 10.1055/a-2465-7283
ISSN 2364-3722

Georg Thieme Verlag KG, Oswald-Hesse-Straße 50, 70469 Stuttgart, Germany

Corresponding author
Dr. Samir C Grover, University of Toronto, Department of Medicine, Toronto, Canada
Samir.grover@utoronto.ca

Supplementary Material is available at https://doi.org/10.1055/a-2465-7283

ABSTRACT

Background and study aims Competent endoscopic ultrasound (EUS) performance requires a combination of technical, cognitive, and non-technical skills. Direct observation assessment tools can be employed to enhance learning and ascertain clinical competence; however, there is a need to systematically evaluate validity evidence supporting their use. We aimed to evaluate the validity evidence of competency assessment tools for EUS and examine their educational utility.

Methods We systematically searched five databases and gray literature for studies investigating EUS competency assessment tools from inception to May 2023. Data on validity evidence across five domains (content, response process, internal structure, relations to other variables, and consequences) were extracted and graded (maximum score 15). We evaluated educational utility using the Accreditation Council for Graduate Medical Education framework and methodological quality using the Medical Education Research Quality Instrument (MERSQI).

Results From 2081 records, we identified five EUS assessment tools from 10 studies. All tools are formative assessments intended to guide learning, with four employed in clinical settings. Validity evidence scores ranged from 3 to 12. The EUS and ERCP Skills Assessment Tool (TEESAT), Glo-

‡ These authors contributed equally.

bal Assessment of Performance and Skills in EUS (GAPS-EUS), and the EUS Assessment Tool (EUSAT) had the strongest validity evidence with scores of 12, 10, and 10, respectively. Overall educational utility was high given ease of tool use. MERSQI scores ranged from 9.5 to 12 (maximum score 13.5).

Conclusions The TEESAT, GAPS-EUS, and EUSAT demonstrate strong validity evidence for formative assessment of EUS and are easily implemented in educational settings to monitor progress and support learning.

Background and study aims

Endoscopic ultrasound (EUS) encompasses a range of diagnostic and therapeutic procedures [1,2]. Competent performance of EUS requires effective endoscopic manipulation and a knowledge of relevant anatomy, typically necessitating additional training beyond core endoscopy training. As the scope of procedures in EUS continues to expand, along with the complexity of skills required and the potential for serious adverse events (AEs) [3], there is a growing need for formal assessment using tools with robust validity evidence. Such assessments are crucial to support learning and ensure that endoscopists attain and maintain competence in performing EUS procedures.

Training in EUS has traditionally occurred in an apprenticeship model, whereby learners acquire technical, cognitive, and non-technical skills under the supervision of experienced faculty. Recently, there have been international efforts to standardize training content [4,5,6]. Similarly, there is a need to standardize competency assessment. Competence in performing EUS has primarily been determined by procedure volume and, more recently, key performance indicators (KPIs) such as visualization of anatomical landmarks and positive fine-needle aspiration (FNA) or biopsy (FNB) rates [7,8]. These approaches, however, have limitations. The procedure volume approach is imperfect because trainees achieve competence at different rates [9]. Likewise, KPIs do not offer meaningful feedback and can be challenging to measure when trainees are performing supervised rather than independent procedures. EUS direct observation competency assessment tools are advantageous and can complement the aforementioned metrics by providing an assessment of the breadth of technical, cognitive, and non-technical skills needed to perform the procedure [6,10,11,12]. Such tools also enable provision of feedback tailored to trainee performance, and ultimately, can be used to inform decisions about competency [10]. To be able to broadly implement use of such assessments, these tools, and the scores they produce, must be well supported by validity evidence.

Messick's unified theory of validity has been endorsed by the American Educational Research Association and the National Council on Measurement in Education as the basis of a framework evaluating educational assessment tool validity [13,14]. Previous systematic reviews have employed this framework in evaluating assessment tools for colonoscopy and endoscopic retrograde cholangiopancreatography (ERCP) [15,16]. In the current systematic review, we aimed to use Messick's framework to evaluate validity evidence supporting direct observation competency assessment tools for EUS and examine the educational utility of each assessment instrument.

Methods

We performed a systematic review according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement. It was registered a priori on the International Prospective Register of Systematic Reviews (PROSPERO) (CRD42023427277).

Study selection

We aimed to identify studies examining validity evidence for procedure-specific tools that assess competency in performing EUS by direct observation. Included tools could be employed in a clinical or simulated setting and utilized by the participant themselves (self-assessment) or by a rater (external assessment). Studies were excluded if: (1) they assessed endoscopic procedures other than EUS (e.g., ERCP, colonoscopy, esophagogastroduodenoscopy); (2) they employed a direct observation assessment tool but did not discuss validity evidence, determined a priori as described below; (3) they provided only simulator outputs (e.g., procedure time) or assessed solely the cytopathology/diagnostic yield of EUS-FNA; (4) they did not include any gastroenterologists and/or surgeons; and (5) the proposed direct observation assessment tool was not utilized in the study.

Search strategy

Five databases were searched from inception until May 18, 2023: MEDLINE, EMBASE, and Evidence Based Medicine Reviews, which includes Cochrane Database of Systematic Reviews, Cochrane Central Register of Controlled Trials, and ACP Journal Club (full details of search strategy and results are provided in **Supplementary Table 1** and **Supplementary Table 2**, respectively). We further reviewed reference lists of included studies, relevant existing reviews, as well as conference proceedings from Digestive Diseases Week and United European Gastroenterology Week from 2011 to 2023. Two authors (AC and HH) screened titles, abstracts, and full texts for eligibility, independently. Disagreements were resolved by two other authors (RK and SCG) through consensus. Covidence (Veritas Health Information, Melbourne, Australia) was used to import and manage reference, abstract, and full-text review.

Data extraction

Two authors (AC and HH), independently and in duplicate, extracted data into a standardized sheet, which included details regarding the assessment purpose, setting, tool format (e. g., Likert scale), and competency domains (technical, cognitive, and/or non-technical). Information regarding the assessors and trainees, validity evidence, educational utility, and methodological quality, was also collected.

Validity evidence

Messick's unified framework was used as the basis to define validity evidence a priori [14]. It consists of five domains: content, which comprises the steps taken to ensure the tool's content is relevant and representative; response process, evaluating the actions of the assessors and trainees and how they relate to the proposed construct; internal structure, which examines how the tool's items relate to each other and the intended construct; relations to other variables, which regards the correlation between the tool's outputs and scores with other measures; and consequences, evaluating the short- and long-term effects of the score interpretation, both intended and unintended [17]. **Supplementary Table 3** provides more detail on the validity framework.

Two authors (AC and HH) independently assessed the five aforementioned sources of validity evidence for each assessment tool [18]. Disagreements were resolved by a third author (RK) as needed. Possible scores were 0 (no discussion regarding validity evidence), 1 (limited discussion regarding validity evidence), 2 (supportive but limited information regarding validity evidence and the significance of scoring), and 3 (strongly supportive information regarding validity evidence and the significance of scoring). The maximum total score was 15. Members of the study team with expertise in endoscopic assessment have previously employed this scoring framework to evaluate validity evidence supporting assessment tools for both colonoscopy and ERCP (RK, CMW, SCG).

Educational utility

Educational utility was assessed using the Accreditation Council for Graduate Medical Education (ACGME) standards [19]. Four areas were evaluated: ease of use, ease of interpretation, resources required, and educational impact. As detailed in Supplementary Table 3, educational impact was graded as A (meets both standards 1 and 2), B (meets standards 1 or 2), C (meets standard 3), or N (not enough information to judge).

Methodological quality

Study quality was assessed using the Medical Education Research Quality Instrument (MERSQI). MERSQI has been shown to be reliable for assessing research in medical education, with at least substantial interrater reliability on each of its eight items [20]. We excluded two items; the item assessing validity evidence for evaluation instrument scores, because this was completed comprehensively as the purpose of our study, and the item assessing sampling response rate, given study participants did not respond to surveys. The maximum score for the remaining six items was 13.5.

Outcomes

The primary outcome of this study was the strength of validity evidence supporting the identified direct observation competency assessment tools for EUS. The secondary outcome was the educational utility of each assessment tool. Evidence from multiple individual studies was pooled to evaluate the primary and secondary outcomes to form an overall opinion of the assessment tool.

Statistical analysis

Inter-rater agreement for validity evidence and educational utility was evaluated using Cohen's kappa (k), ranging from poor (< 0), slight (0 to < 0.2), fair (0.2 to < 0.4), moderate (0.4 to < 0.6), substantial (0.6 to < 0.8), and almost perfect (0.8 to 1). Statistical Package for the Social Sciences (SPSS) was used for statistical analysis (Version 24.0, IBM Corporation, Armonk, New York, United States). We did not conduct quantitative analyses given the substantial heterogeneity in study design and score reporting.

Results

Our electronic database search and gray literature search yielded 2097 records and 11 records, respectively. After de-duplication, 2081 records were screened, of which 44 underwent full-text review. Ten studies covering five EUS assessment tools were included [9, 21, 22, 23, 24, 25, 26, 27, 28, 29] (**Supplementary Fig. 1**). Inter-rater agreement at the full-text screening stage was moderate (k = 0.53), with a proportionate agreement of 82%.

EUS assessment tools

All five identified EUS assessment tools were designed for formative assessment (i. e., low stakes, assessment for learning) (▶ **Table 1**). Each tool was rater-based and completed interprocedurally. The Global Assessment of Performance and Skills in EUS (GAPS-EUS) [23, 24] and The EUS and ERCP Skills Assessment Tool (TEESAT) [26, 27] had additional post-procedure self-assessment components. All five tools were applied in a live assessment setting with the Endoscopic Ultrasound Assessment Tool (EUSAT) having an additional video component that was reviewed post-procedure [21]. Four of the tools were employed in a clinical setting [9, 21, 23, 24, 25, 26, 27, 28, 29], with the fifth employed in a simulated setting on live pigs [22]. All tools assessed technical competency, with the GAPS-EUS and TEESAT also examining cognitive and non-technical skills in EUS [9, 21, 22, 23, 24, 25, 26, 27, 28, 29]. All five tools assessed visualization/recognition of anatomical landmarks [9, 21, 22, 23, 24, 25, 26, 27, 28, 29]. The EUSAT assessed solely mediastinal anatomical landmarks, whereas the other four tools also assessed pancreatic and hepatobiliary landmarks [9, 21, 22, 23, 24, 25, 26, 27, 28, 29].Three tools assessed biopsy sampling technique [9, 21, 23, 24, 25, 26, 27, 28] and two assessed the trainee's ability to describe aspects of the pathology [9,

▶ **Table 1** Characteristics of EUS assessment tools.

Tool [reference (s)]	Assessment purpose	Setting and procedures	Endoscopist population	Type of scale	Competency domains assessed	Assessment method
3-point Likert Scale [22]	Formative	Simulated (live pig model) Pancreaticobiliary EUS	Physicians: Fellowship trainees	5-item checklist; 3-point Likert scale assessing visualization of anatomical structure	Technical	Live external assessor
EUSAT [21]	Formative	Clinical Mediastinal EUS*	Physicians: 3 EUS trainees and 3 experienced physicians in Pulmonary Medicine and Gastroenterology	12-item checklist; 5-point Likert Scale assessing insertion of endoscope, knowledge of anatomical landmarks, and biopsy sampling	Technical	Live and video-based external assessor
GAPS-EUS (2021) [23, 24] 00.00.0000 00:00:00	Formative	Clinical Mediastinal, pancreaticobiliary, and luminal EUS	Physicians: EUS fellows in training	5-item checklist; 5-point Likert scale	Technical [23, 24], cognitive [23, 24], non-technical [24])	Live external assessor and
Point-score system [29]	Formative	Clinical Mediastinal, pancreaticobiliary, and luminal EUS	4 physicians: gastroenterology fellows 1 nurse endoscopist	Point score system; Likert scale with points given for ability to produce high quality view with certainty	Technical	Live external assessor
TEESAT [9, 25, 26, 27, 28] 0/0/0000 0:00:00 AM	Formative	Clinical Pancreaticobiliary and luminal EUS (with one mediastinal station)	Physicians: Advanced endoscopy trainees and independent staff	1)18-item checklist; 4-point [26, 27, 28] and 5-point Likert scales assessing technical and cognitive aspects [9, 25] 2) global rating scale; 10-point system [26] and 4-point system [27, 28] 3)7-item checklist with self-assessment of comfort performing the procedure [26, 27]	Technical, cognitive, and non-technical	Live external assessor and self-assessment

EUS, endoscopic ultrasound; EUSAT, Endoscopic Ultrasound Assessment Tool; GAP-EUS, Global Assessment of Performance and Skills in EUS; TEESAT, The EUS and ERCP skills assessment tool.
*The endoscopic ultrasound assessment tool (EUSAT) also included examination of the liver.

23, 24, 25, 26, 27, 28]. The GAPS-EUS also assessed patient management [23, 24], whereas the TEESAT assessed management planning [26, 27, 28] and procedure complications [9, 25, 26, 27, 28].

All tools employed a Likert-type rating scale, with the anchors of two tools being based on the level of assistance required [9, 23, 24, 25, 26, 27, 28] and the other three tools' anchors based on quality of anatomical visualization/technique [21, 22, 29]. The TEESAT [26, 27, 28] also employed a Likert scale for global rating items assessing overall technical and cognitive competence in EUS. In all 10 studies, the respective tools were used to assess gastroenterology and/or advanced endoscopy trainees performing EUS on adult patients [9, 21, 22, 23, 24, 25, 26, 27, 28, 29]. Meenan et al. also assessed a nurse endoscopist and Konge et al. assessed experienced staff physicians, some of whom were respirologists [21, 29].

Validity evidence

Inter-rater agreement for validity evidence was substantial (k = 0.73, raw agreement 80%). Overall scores ranged from 3 [22] to 12 [9, 25, 26, 27, 28] (▶ **Table 2**, ▶ **Fig. 1**, and **Supplementary Table 4**). Three assessment tools were piloted in earlier studies, informing changes for use in later studies [9, 21, 23, 24, 25, 26, 27, 28]. Response process validity evidence was available for four tools, [9, 22, 23, 24, 25, 26, 27, 28]. Internal structure validity evidence was described for three tools, including a critical analysis of the data distribution for three instruments [9, 21, 23, 24, 25, 26, 27, 28] and inter-rater reliability was reported for two tools [21, 23, 24]. All five tools had evidence of relations

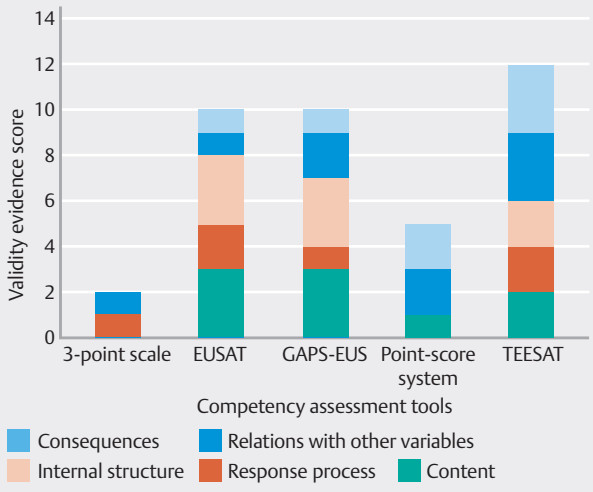
▶ **Table 2** Comparison of three EUS assessment tools with strong validity evidence.

Validity evidence domain	Assessment tool		
	TEESAT [9, 25, 26, 27, 28]	**GAPS-EUS [23, 24]**	**EUSAT [21]**
Content	▪ Pilot testing and revision: Y; Tool was revised throughout multiple multi-center studies ▪ Clinical guidelines: Y; Quality indicators in EUS were reviewed when creating the tool	▪ Pilot testing and revision: Y; Hedenström et al. (2015) collected data between May-November 2014 at one center while the study by Hedenström et al. (2021) collected data between 2014–2019 at two centers with revisions made to the tool [23, 24] ▪ Clinical guidelines: Y; Based on performance measures and quality indicators recommended by the American Society for Gastrointestinal Endoscopy and the European Society of Gastrointestinal Endoscopy.	▪ Expert panel: Y; Tool developed by experts in gastroenterology, pulmonary medicine, and educational measurement. ▪ Pilot testing and revision: Y; First draft of assessment form revised after use in pilot study. ▪ Scoring framework: Y; Based on anatomical knowledge of mediastinal structures, ability to take biopsies and maneuver scope. ▪ Other: Y; Participants were given written and verbal information on items to be assessed.
Response process	▪ Test security: Y; Data were entered and analyzed from a centralized national database [26, 27, 28]. Trainees were de-identified and data were entered into an unspecified database [25]. ▪ Other: Y; Each trainee was graded on every 3rd[24], 5th[27, 28], and 10th[9, 25]EUS procedure, as well as after the 25th EUS with or without FNA was completed. To reduce halo effect, recall bias, and recency effect, grading was done immediately following the procedure [26, 27, 28]. By continuously studying the control charts, the performance of each trainee was compared with a predetermined standard, allowing for the detection of negative trends and enabling earlier feedback (retraining or continued observation) [28].	▪ Rater data analysis: Y; Observer and trainee scores indicated with 95% CI interval ▪ Rationale for composite outcomes: Y; Overall trainee, observer, and compound score formulas provided and justified. Justification provided for scoring EUS-FNA separately [24].	▪ Test security: Y; Recording sent through hard drive to one rater ▪ Effect of rater training: Y; Study included an independent, international rater who did not participate in the development of the assessment tool, did not know the training level of the participants, and was only provided with written instructions about the use of the tool ▪ Other: Y; Recording was standardized: all started and ended with the same EUS step. Video recordings were assessed by two experts blindly, and independently. Additionally, a third assessor reviewed all recordings without consulting other assessors
Internal Structure	▪ Critical analysis of data distribution: Y; No significant change in the proportion of trainees achieving competency was noted when varying the definition of competency to become more stringent (> 1, no assistance [25] and < 5% failure rate [26]. ▪ Other: Y; Wani et al. (2017) compared the basic attributes (number of trainees/year, annual volume of EUS offered during training) between participating and non-participating programs, and no differences were noted between the 2 groups [26].	▪ Internal consistency: Y; Observer version of the tool (Cronbach's α = 0.87) and trainee version of the tool (Cronbach's α = 0.89)[24]. ▪ Inter-rater reliability: Y; High - Correlation coefficient between observer and trainee score: (r = 0.83, r2 = 0.69, P < 0.001) [24]. Good correlation between overall mean score of observer and performer (r = 0.66, P < 0.001) [23]. ▪ Other reliability: Y; Bland-Altman plot of the observer and trainee score in all trainee performed EUS procedures: similar distribution among low and high score procedures [24]. ▪ Critical analysis of data distribution: Y; Trainees overestimated their own performances in comparison to assessors - Dunning Kruger effect [24]. Higher observed GEUSP score (3.3 compared to 2.9, P = 0.09) for patients enrolled later (n = 13) than patients enrolled first (n = 12) with one performer [23].	▪ Inter-rater reliability: Y; Cronbach's α (as a comparison of scores of all raters based on blinded video recordings) = 0.93 ▪ Intra-rater reliability: Y; There is strong correlation between scoring recorded by direct observation and that from blinded video-recordings by the same rater, which gave a good intra-rater reliability (Cronbach's α = 0.80) ▪ Test-retest reliability: Y; Evaluation was done under direct observation and after 2 months from video recordings: Cronbach's α = 0.80 ▪ Critical analysis of data distribution: Y; Trainee scores were lower under direct observation and consultants scored higher with no blinding. Generalizability analysis showed that a single rater assessing a single EUS procedure would only achieve a generalizability coefficient of 0.47, and assessment of a single procedure would still be unsure even when using three raters (generalizability coefficient = 0.53)

▶ **Table 2** (Continuation)

Validity evidence domain	Assessment tool		
	TEESAT [9, 25, 26, 27, 28]	**GAPS-EUS [23, 24]**	**EUSAT [21]**
Relations with other variables	▪ Learner characteristic, general training: Y; No difference noted between trainees with experience and without in proportion of those achieving competence (P = 0.99) [27]. ▪ Learner characteristic, other: Y; Findings were discordant with number of EUS procedures performed when estimated via a trainer "global assessment" of competency (165 EUS examinations) [28]. ▪ Other: Y; Overall agreement between results obtained by using the global rating scale and those using TEESAT was fair for competence in EUS (overall technical: k = 0.38 [95% CI, 0–0.79] [26]. Agreement between TEESAT and the global rating scale for EUS competence was fair (technical: k = 0.36, 95% CI 0.02–0.74; cognitive: k = 0.36, 95% CI 0.01–0.74) [27].	▪ Learner characteristic, general training: Y; Correlation between overall mean GEUSP score and previous overall EUS experience: r = 0.9, P = 0.06 [23]. ▪ Separate measure, patient care: Y; Patient adverse event rate 2/157 [24].	▪ Learner characteristic, general training: Y; Mean score of experienced physicians (40.6) was significantly higher than mean score of trainees (31.5)(P = 0.001). The scores of the experienced physicians were also less variable (SD = 3.3 vs. 8.0; P = 0.00). ▪ Other: Y; Trainees scored approximately 10% lower when the rater knew their identity.
Consequences	▪ Rigorous pass/fail cut-point, established approach: Y; The number of EUS procedures with FNA required for an average trainee to achieve competence in EUS-FNA was 110 (95% CI, 90–140); at this time point, the average trainee had completed 226 EUS examinations [28]. A specific case load does not ensure competency in EUS - 225 cases should be considered the minimum caseload for training as no trainee achieved competency before this point [25]. ▪ Evaluation of actual pass rate: Y; Acceptable rate of 10% and unacceptable failure rate of 30% showed similar results to the corresponding rates of 10% and 20%, respectively [25]. The vast majority of AETs achieved competence in overall cognitive (76.4%) and overall technical (82.3%) aspects of EUS at the end of their training [24], 91.7% achieved both cognitive and technical success [27]. ▪ Anticipated impact: Y; The results of this study demonstrate the substantial variability in the learning curves and number of AETs achieving competence in EUS (overall and individual endpoints) [26, 27]. ▪ Unanticipated impact: Y; Post-study questionnaire showed that there is a lack of concordance between the results of competence as assessed by learning curve analysis and comfort level expressed by AETs in independently performing EUS after completion of their advanced endoscopy training [27]. Data suggest that endoscopy trainers may overestimate competence using global assessment of competence and trainees may benefit from a forced evaluation of their trainees' individual core skills [27].	▪ Rigorous pass/fail cut-point, unproven approach: Y; An Observer Score of at least four in multiple, consecutive GAPS-EUS assessments could be a reasonable threshold as the cut-off level for sufficiently high competence in EUS [24].	▪ Anticipated impact: Y; The EUSAT can be used as a quality control measure: EUS procedure recordings can be sent to international experts for review. EUSAT assessment framework can be utilized in settings where consistent supervision by EUS experts is not feasible.

CI, confidence interval; EUS, endoscopic ultrasound; EUSAT, Endoscopic Ultrasound Assessment Tool; GAPS-EUS, Global Assessment of Performance and Skills in EUS; TEESAT, The EUS and ERCP Skills Assessment Tool (TEESAT);
Note: Statistical data presented in Table 2 are from the original publications and pertain to the five sources of validity evidence assessed.



▶ **Fig. 1** EUS competency assessment tool validity evidence scores.

with other variables related to participant characteristics, including level of training [9, 21, 23, 24, 25, 26, 27, 28, 29] or EUS experience [22, 29]. Data supporting evidence of consequences were present for four tools, with only the TEESAT and Point-Score system outlining rigorous pass/fail cut-points [9, 25, 26, 27, 28, 29] and the GAPS-EUS proposing an unsupported pass/fail cut-point [23, 24].

The EUSAT, GAPS-EUS, and TEESAT had the strongest validity evidence, with overall scores of 10, 10, and 12, respectively [9, 21, 23, 24, 25, 26, 27, 28]. The EUSAT and GAPS-EUS both had strongly supportive data for content and internal structure validity evidence and weakly supportive evidence for the consequences domain [21, 23, 24]. The GAPS-EUS [23, 24] had intermediate evidence for relations with other variables, while the EUSAT had weak evidence in this domain [21]. The TEESAT had strongly supportive data in the relations with other variables and consequences domains and intermediate evidence with respect to the content, response process, and internal structure domains [9, 25, 26, 27, 28].

Educational utility

Inter-rater agreement for educational utility was very high (k = 0.92, raw agreement 96%). All five tools were graded as an A for ease of use, given they all require no set-up and are easily accessible (▶ **Table 3**). All five tools received a B for ease of interpretation because none had normative data available but had easily interpretable scores. All five tools received an A for resources required because they did not require any additional resources beyond the documentation tool and no training was needed to use the tool. Finally, the TEESAT received an A for educational impact because it positively impacted credentialing and competence thresholds by proposing minimum standards for case volume exposure during training. In contrast, the other tools received a B rating [21, 29] or did not comment on educational utility [22, 23, 24].

Study quality

MERSQI scores for the studies evaluated ranged from 9.5 [22] to 12 [24, 27], of a maximum possible score of 13.5 (▶ **Table 4**). All studies were observational [9, 21, 22, 23, 24, 25, 26, 27, 28]. Five studies were conducted at more than two institutions [9, 25, 26, 27, 28] and only two studies reported patient AEs [24, 27].

Discussion

In this systematic review, we appraised the validity evidence of direct observation competency assessment tools for EUS. We included 10 studies which provided validity evidence for five tools: 3-point Likert Scale [22], EUSAT [21], GAPS-EUS [23, 24], Point-score system [29], and TEESAT [9, 25, 26, 27, 28]. All tools evaluated technical competency in performing EUS through direct observation by raters and were designed for formative assessment purposes. Applying Messick's unified validity framework, the TEESAT, GAPS-EUS, and EUSAT [9, 21, 23, 24, 25, 26, 27, 28] demonstrated the strongest validity evidence. With respect to educational utility, all tools were easy to use and required minimal resources for implementation. However, none of the tools included normative comparison data. Discussion

▶ **Table 3** Educational utility of EUS assessment tools.

Tool [reference(s)]	Educational utility*			
	Ease of use	Ease of interpretation	Resources required	Educational impact
3-point Likert Scale [22]	A	B	A	N
EUSAT [21]	A	B	A	B
GAPS-EUS [23, 24]	A	B	A	N
Point-Score System [29]	A	B	A	B
TEESAT [9, 25, 26, 27, 28]	A	B	A	A

EUSAT, Endoscopic Ultrasound Assessment Tool; GAPS-EUS, Global Assessment of Performance and Skills in EUS (GAPS-EUS); TEESAT, The EUS and ERCP Skills Assessment Tool.
*Possible ratings were: A (meets standards 1 and 2), B (meets standard 1 or 2), C (meets standard 3), or N (not enough information to judge). See Supplementary Table 3. Adapted from the 2009 Accreditation Council for Graduate Medical Education report for evaluating assessment tools [19].

▶ **Table 4** Methodological quality of studies that provided validity evidence for EUS assessment tools, rated using the Medical Education Research Quality Instrument (MERSQI).

Tool [reference(s)]	Methodologic quality*						
	Study design† (max 3)	Institutions‡ (max 1.5)	Type of data§ (max 3)	Data analysis sophistication¶ (max 2)	Data analysis appropriatenes†† (max 1)	Highest outcome type‡‡ (max 3)	Total Score (max 13.5)
3-point Likert Scale							
Barthet et al. (2007) [22]	1.5	0.5	3	2	1	1.5	9.5
EUSAT							
Konge et al. (2012) [21]	1	1	3	2	1	2	10
GAPS-EUS							
Hedenström et al. (2015) [23]	1	0.5	3	2	1	3	10.5
Hedenström et al. (2021) [24]	2	1	3	2	1	3	12
Point-Score System							
Meenan et al. (2003) [29]	1.5	0.5	3	2	1	2	10
TEESAT							
Wani et al. (2013) [9]	1.5	1.5	3	2	1	2	11
Wani et al. (2015) [25]	1.5	1.5	3	2	1	2	11
Wani et al. (2017) [26]	1.5	1.5	3	2	1	2	11
Wani et al. (2018) [27]	1.5	1.5	3	2	1	3	12
Wani et al. (2019) [28]	1.5	1.5	3	2	1	2	11

EUSAT, Endoscopic Ultrasound Assessment Tool; GAPS-EUS, Global Assessment of Performance and Skills in EUS; TEESAT, The EUS and ERCP Skills Assessment Tool (TEESAT).
*Methodological quality judged using the Medical Education Research Quality Instrument (MERSQI) [20].
†Study design: 1 = single-group cross-sectional or single-group post-test only; 1.5 = single-group pretest and post-test; 2 = nonrandomized > 1 group; 3 = randomized controlled trial.
‡Institutions: 0.5 = one institution; 1 = two institutions; 1.5 = more than two institutions.
§Type of data: 1 = assessment by study participant; 3 = objective.
¶ Data analysis sophistication: 1 = descriptive analysis only; 2 = beyond descriptive analysis.
††Data analysis appropriateness: 1 = data analysis appropriate for study design.
‡‡Highest outcome type: 1.5 = knowledge or skills in simulated setting; 2 = performance in clinical setting; 3 = patient/healthcare outcome.

of educational utility varied among studies, with the TEESAT's educational impact being described most extensively [9, 25, 26, 27, 28]. Most studies were of high methodological quality based on study design, but a notable difference was that only specific TEESAT and GAPS-EUS studies reported on patient outcomes [24, 27].

The strength of validity evidence for the five identified EUS direct observation competency assessment tools varied significantly across the domains of content, response process, internal structure, relations with other variables, and consequences validity. The 3-point Likert scale [22] and Point-score system [29] had relatively weaker validity evidence, whereas the TEESAT, GAPS-EUS, and EUSAT had relatively stronger validity evidence [9, 21, 23, 24, 25, 26, 27, 28]. In Barthet et al., the 3-point Likert scale was not the primary focus of the paper. Rather, the study centered on creation of a new EUS training model and reported on trainee performance [22]. Meenan et al. discussed three of the five elements of the validity framework in relation to the Point-Score system [29]. The tool was used to propose competency cut-offs and compare performance of a nurse endoscopist to physicians in radial EUS, making it the only study to include a non-physician participant. However, there were no data describing the tool's response process or internal structure. Overall, there is insufficient validity evidence to recommend use of either the 3-point Likert scale or the Point-Score system tool in educational or credentialing settings.

Studies describing the TEESAT, GAPS-EUS, and EUSAT demonstrated evidence for all five components of the validity framework [9, 21, 23, 24, 25, 26, 27, 28]. Furthermore, these tools were all piloted and revised in subsequent studies. The EUSAT, which focuses on technical skills, was the only tool evaluated in both the live context (unblinded raters) and video assessment

context (blinded raters) [21]. This assessment tool demonstrated strong internal structure validity evidence. Generalizability theory was employed to determine that a generalizability coefficient of 0.7, suitable for low-stakes formative assessment (e.g. in-training assessments), could be achieved with one rater assessing three procedures. Meanwhile, a generalizability coefficient of 0.8, suitable for higher-stakes assessments, could be achieved with one rater assessing five procedures or two raters each assessing four procedures. Of note, validity evidence for the EUSAT was examined in the context of gastroenterologists and pulmonary medicine specialists performing mediastinal examinations for staging of non-small cell lung cancer, thus limiting the tool's applicability to other EUS procedures (e.g., pancreatic, biliary, hepatic). Evidence in a gastrointestinal context is required prior to use for assessment of gastroenterology trainees or faculty [21].

The GAPS-EUS and TEESAT exhibit similarities [9, 23, 24, 25, 26, 27, 28]. Both tools' Likert-scale options are based on the level of assistance participants require. Furthermore, they are the only published EUS assessment tools that asses all three competency domains: technical skills (e.g., scope manipulation, anatomic visualization, achievement of FNA), cognitive skills (e.g. identification of pathology, management decision-making), and non-technical skills (e.g., communication). They also both assess aspects of patient care, with the GAPS-EUS assessing patient management and the TEESAT capturing procedure complications. As a result, both assessment tools provide a comprehensive assessment of trainee performance, aligning with established pre-procedure, intra-procedure, and post-procedure EUS quality indicators [30].

Distinctions between the GAPS-EUS and TEESAT lie in the differing components of the validity framework upon which the studies evaluating the tools focus. The GAPS-EUS is notable for its strong internal structure [23, 24], with almost perfect internal consistency in both its external rater and self-rater versions. There was also a high correlation between rater-assessed and self-assessed scores. However, a critical analysis of the data suggested a Dunning Kruger effect, wherein trainees tended to overestimate their performance compared with an external assessor's rating. In contrast, the TEESAT does not have strong data supporting internal structure validity evidence. However, it has robust evidence of consequences validity, an area where the GAPS-EUS lacks evidence. Using learning curve data from 32 centers, Wani et al. in 2019 established that 225 EUS cases (including 100 FNA) are required, on average, for advanced endoscopy trainees to achieve competence in EUS, providing data to help establish minimum standards for case volume exposure during training [28]. Furthermore, they demonstrated wide variability in EUS trainee learning curves and found differential learning curves for technical and cognitive skills. Finally, they noted that global assessment scales may overestimate trainee skills when used in isolation.

Overall, there are compelling reasons to implement direct observation competency assessment tools with strong validity evidence for use in assessing endoscopists performing EUS. In late 2023, the European Society of Gastrointestinal Endoscopy (ESGE) released a position statement outlining an EUS training curriculum framework, with competency assessment using valid tools being one of their main recommendations [4]. Assessment is known to drive learning by offering a structured approach to identifying areas for improvement and guiding skill development. By utilizing procedure-specific assessment tools, educators can ensure that trainees receive targeted feedback, allowing them to focus on specific areas of their performance that may need enhancement [10, 31]. In EUS, Hedenström et al.'s confirmation of the Dunning-Kruger effect, wherein trainees overestimate their skill level, highlights the importance of external assessments in providing a more objective measure of competency [24]. Wani et al.'s finding of differential learning curves for technical and cognitive skills underscores the importance of selecting assessment tools that cover the breadth of skills necessary for competent EUS performance [26, 27]. Such holistic tools were also shown to provide more value compared with global ratings, which tended to overestimate skill in performing EUS. Furthermore, implementing these tools is straightforward, because they require no additional resources or training, reducing barriers to implementation.

Although there are numerous benefits to incorporating assessment tools into training, existing literature also underscores potential limitations. For instance, a study analyzed the performance of endoscopists during their first year of independent practice and found no relationship between achieving competence at the end of training, as determined by TEESAT scores, and performance on key EUS quality indicators in practice [27]. This finding raises doubt about the predictive ability of direct observation assessment tools regarding future EUS performance, although the study had relatively few participants. Longer-term and larger datasets are needed to confirm or refute the assertion that direct observation assessment tools can be used to predict clinically relevant outcomes.

This study has several limitations. First, although the study was conducted systematically utilizing an established validity framework, with two independent reviewers and an experienced team in medical education, assessment of validity evidence, while structured, is inherently subjective. Second, only studies that were formally published and in English were included, potentially excluding relevant evidence. Last, most studies did not specifically comment on the educational relevance of their findings, potentially limiting generalizability. Strengths of this review include a comprehensive search strategy, developed in conjunction with a health sciences librarian, and inclusion of five databases to capture relevant citations. We employed specific criteria to quantify the validity evidence and educational utility of included instruments as well as the methodological rigor of included studies, achieving excellent inter-rater reliability.

Conclusions

In conclusion, as use of EUS continues to expand, it is imperative for training programs to integrate rigorously developed direct observation competency assessment tools with robust validity evidence into their curricula. These tools are essential in supporting feedback provision, fostering learner development,

and ensuring achievement of competence. Overall, we found that the TEESAT, GAPS-EUS, and EUSAT have the strongest validity evidence for formative assessment and are easy to implement in educational settings. Although the TEESAT and GAPS-EUS are broadly applicable, validity evidence for the EUSAT is currently limited to mediastinal examinations. Future research should focus on identifying and addressing barriers to implementation and evaluating the utility of these tools in summative,

Conflict of Interest

N. Forbes does not have any current conflicts of interest but in the last 3 years was a consultant for AstraZeneca, a consultant and speaker for Boston Scientific, and a consultant and speaker for Pentax Medical. He received research funding from Pentax Medical. J. D. Mosko has been a consultant and speaker for Boston Scientific, ERBE, Fuji, Medtronic, Pendopharm, and Steris. He received research funding from Boston Scientific and ERBE. S. C. Grover has been a speaker for Abbvie, a stockholder and employee of Volo Healthcare, on the advisory board for Amgen, BioJAMP, Pfizer, and Sanofi, and has received education support from Fresenius Kabi, BioJAMP, Celltrion, Takeda, and Pfizer. A. Ceccacci, H. Hothi, R. Khan, N. Gimpaya, B. Chan, P. D. James, D. J. Low, E. Yeung, and C. M. Walsh do not have any conflicts of interest to indicate.

References

[1] Levine I, Trindade AJ. Endoscopic ultrasound fine needle aspiration vs fine needle biopsy for pancreatic masses, subepithelial lesions, and lymph nodes. World J Gastroenterol 2021; 27: 4194–4207 doi:10.3748/wjg.v27.i26.4194

[2] Friedberg SR, Lachter J. Endoscopic ultrasound: Current roles and future directions. World J Gastrointest Endosc 2017; 9: 499–505 doi:10.4253/wjge.v9.i10.499

[3] ASGE Standards of Practice Committee. Forbes N, Coelho-Prabhu N et al. Adverse events associated with EUS and EUS-guided procedures. Gastrointest Endosc 2022; 95: 16–26.e2

[4] Karstensen JG, Nayahangan LJ, Konge L et al. A core curriculum for basic EUS skills: An international consensus using the Delphi methodology. Endosc Ultrasound 2022; 11: 122–132 doi:10.4103/EUS-D-21-00125

[5] Cassani L, Aihara H, Anand GS et al. Core curriculum for EUS. Gastrointest Endosc 2020; 92: 469–473 doi:10.1016/j.gie.2020.06.054

[6] Badaoui A, Teles de Campos S, Fusaroli P et al. Curriculum for diagnostic endoscopic ultrasound training in Europe: European Society of Gastrointestinal Endoscopy (ESGE) Position Statement. Endoscopy 2023; 56: 222–240

[7] Patel SG, Keswani R, Elta G et al. Status of competency-based medical education in endoscopy training: A nationwide survey of US ACGME-accredited gastroenterology training programs. Am J Gastroenterol 2015; 110: 956–962 doi:10.1038/ajg.2015.24

[8] Johnson G, Webster G, Boškoski I et al. Curriculum for ERCP and endoscopic ultrasound training in Europe: European Society of Gastrointestinal Endoscopy (ESGE) Position Statement. Endoscopy 2021; 53: 1071–1087 doi:10.1055/a-1537-8999

[9] Wani S, Coté GA, Keswani R et al. Learning curves for EUS by using cumulative sum analysis: implications for American Society for Gastrointestinal Endoscopy recommendations for training. Gastrointest Endosc 2013; 77: 558–565 doi:10.1016/j.gie.2012.10.012

[10] Walsh CM. In-training gastrointestinal endoscopy competency assessment tools: Types of tools, validation and impact. Best Pract Res Clin Gastroenterol 2016; 30: 357–374 doi:10.1016/j.bpg.2016.04.001

[11] Shiha MG, Ravindran S, Thomas-Gibson S et al. Importance of non-technical skills: SACRED in advanced endoscopy. Frontline Gastroenterol 2023; 14: 527–529 doi:10.1136/flgastro-2023-102434

[12] Ravindran S, Haycock A, Woolf K et al. Development and impact of an endoscopic non-technical skills (ENTS) behavioural marker system. BMJ Simul Technol Enhanc Learn 2021; 7: 17–25 doi:10.1136/bmjstel-2019-000526

[13] Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med 2006; 119: 166.e7–16 doi:10.1016/j.amjmed.2005.10.036

[14] Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol 1995; 50: 741–749

[15] Khan R, Zheng E, Wani SB et al. Colonoscopy competence assessment tools: a systematic review of validity evidence. Endoscopy 2021; 53: 1235–1245 doi:10.1055/a-1352-7293

[16] Khan R, Homsi H, Gimpaya N et al. Validity evidence for observational ERCP competency assessment tools: a systematic review. Endoscopy 2023; 55: 847–856 doi:10.1055/a-2041-7546

[17] Faigel DO, Baron TH, Lewis B et al. Ensuring competence in endoscopy - prepared by the ASGE Taskforce on Ensuring Competence in Endoscopy. Oak Brook, IL: American College of Gastroenterology Executive and Practice Management Committees. 2006: https://www.asge.org/docs/default-source/education/practice_guidelines/doc-competence.pdf?sfvrsn=1bfd4951_6

[18] Ghaderi I, Manji F, Park YS et al. Technical skills assessment toolbox: a review using the unitary framework of validity. Ann Surg 2015; 261: 251–262 doi:10.1097/SLA.0000000000000520

[19] Swing SR, Clyman SG, Holmboe ES et al. Advancing resident assessment in graduate medical education. J Grad Med Educ 2009; 1: 278–286 doi:10.4300/JGME-D-09-00010.1

[20] Cook DA, Reed DA. Appraising the quality of medical education research methods: the Medical Education Research Study Quality Instrument and the Newcastle-Ottawa Scale-Education. Acad Med J Assoc Am Med Coll 2015; 90: 1067–1076 doi:10.1097/ACM.0000000000000786

[21] Konge L, Vilmann P, Clementsen P et al. 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 doi:10.1055/s-0032-1309892

[22] Barthet M, Gasmi M, Boustiere C et al. EUS training in a live pig model: does it improve echo endoscope hands-on and trainee competence? Endoscopy 2007; 39: 535–539

[23] Hedenström P, Lindkvist B, Sadik R. Global assessment of EUS performance skills (GEUSP) - a new tool and approach to objectify and measure the learning curve and technical skills of endosonographists. Gastrointest Endosc 2015; 81: AB442

[24] Hedenström P, Marasco G, Eusebi LH et al. GAPS-EUS: a new and reliable tool for the assessment of basic skills and performance in EUS among endosonography trainees. BMJ Open Gastroenterol 2021; 8: e000660

[25] Wani S, Hall M, Keswani RN et al. Variation in aptitude of trainees in endoscopic ultrasonography, based on cumulative sum analysis. Clin Gastroenterol Hepatol 2015; 13: 1318–1325.e2

[26] Wani S, Keswani R, Hall M et al. A prospective multicenter study evaluating learning curves and competence in endoscopic ultrasound and endoscopic retrograde cholangiopancreatography among advanced endoscopy trainees: The rapid assessment of trainee endoscopy skills study. Clin Gastroenterol Hepatol 2017; 15: 1758–1767.e11

[27] Wani S, Keswani RN, Han S et al. Competence in endoscopic ultrasound and endoscopic retrograde cholangiopancreatography, from training through independent practice. Gastroenterology 2018; 155: 1483–1494.e7

[28] Wani S, Han S, Simon V et al. Setting minimum standards for training in EUS and ERCP: results from a prospective multicenter study evaluating learning curves and competence among advanced endoscopy trainees. Gastrointest Endosc 2019; 89: 1160–1168.e9

[29] Meenan J, Anderson S, Tsang S et al. Training in radial EUS: what is the best approach and is there a role for the nurse endoscopist? Endoscopy 2003; 35: 1020–1023

[30] Wani S, Wallace MB, Cohen J et al. Quality indicators for EUS. Gastrointest Endosc 2015; 81: 67–80 doi:10.1016/j.gie.2014.07.054

[31] Strandbygaard J, Scheele F, Sørensen JL. Twelve tips for assessing surgical performance and use of technical assessment scales. Med Teach 2017; 39: 32–37 doi:10.1080/0142159X.2016.1231911