Thieme

The Bethesda ERCP Skills Assessment Tool (BESAT) can reliably differentiate endoscopists of different experience levels

OPEN ACCESS

Authors

Kevin Liu[1], B. Joseph Elmunzer[2], Sachin Wani[3], Tiffany Taft[4], Catharine M Walsh[5,6,7], Mustafa A Arain[8], Tyler M. Berzin[9], James Buxbaum[10], Christopher DiMaio[11], Syed M. Abbas Fehmi[12], Neil Gupta[13], Sreenivasa Jonnalagadda[14], Vladimir Kushnir[15], John T. Maple[16], Raman Muthusamy[17], Amit Rastogi[18,19], Janak N Shah[20], Amitabh Chak[21], Ashley Faulx[21], Nauzer Forbes[22], Rajesh N Keswani[23]

Institutions

1 Gastroenterology, Banner - University Medical Center Phoenix, Phoenix, United States
2 Division of Gastroenterology, Medical University of South Carolina, Charleston, United States
3 Gastroenterology, University of Colorado and Veterans Affairs Medical Center, Aurora, United States
4 Division of Gastroenterology, Northwestern University Feinberg School of Medicine, Chicago, United States
5 Division of Gastroenterology, Hepatology, and Nutrition and the Research and Learning Institutes, The Hospital for Sick Children, Toronto, Canada
6 Department of Pediatrics, University of Toronto Faculty of Medicine, Toronto, Canada
7 The Wilson Centre, University of Toronto, Toronto, Canada
8 Center for Interventional Endoscopy, AdventHealth Orlando, Orlando, United States
9 Gastroenterology, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, United States
10 Medicine/Gastroenterology, University of California, San Francisco, San Francisco, United States
11 Gastroenterology, Catholic Health St Francis Hospital & Heart Center, Roslyn, United States
12 Internal Medicine, UC San Diego Health System, San Diego, United States
13 Gastroenterology, Loyola University Medical Center, Chicago, United States
14 Gastroenterology, Saint Luke's Health System, Kansas City, United States
15 Gastroenterology, Washington University, St Louis, United States
16 Internal Medicine, University of Oklahoma, Oklahoma City, United States
17 Vatche and Tamar Manoukian Division of Digestive Diseases, University of California, Los Angeles, United States
18 Gastroenterology, Kansas University Medical Center, Kansas City, United States
19 Gastroenterology, Veterans Affairs Medical Center, Kansas City, United States
20 Gastroenterology, Ochsner Medical Center - New Orleans, New Orleans, United States
21 Gastroenterology, UH Cleveland Medical Center, Cleveland, United States
22 Medicine, University of Calgary, Calgary, Canada
23 Medicine, Northwestern University Feinberg School of Medicine, Chicago, United States

Keywords
Pancreatobiliary (ERCP/PTCD), ERC topics, Training, Quality and logistical aspects, Quality management

received 20.3.2023
accepted after revision 22.8.2023
accepted manuscript online 28.8.2023

Bibliography
Endosc Int Open 2024; 12: E324–E331
DOI 10.1055/a-2161-1982
ISSN 2364-3722
© 2024. The Author(s).
This is an open access article published by Thieme under the terms of the Creative Commons Attribution-NonDerivative-NonCommercial License, permitting copying and reproduction so long as the original work is given appropriate credit. Contents may not be used for commercial purposes, or adapted, remixed, transformed or built upon. (https://creativecommons.org/licenses/by-nc-nd/4.0/)
Georg Thieme Verlag KG, Rüdigerstraße 14,
70469 Stuttgart, Germany

Corresponding author
Dr. Kevin Liu, Banner - University Medical Center Phoenix, Gastroenterology, 1441 N 12th St, Phoenix , 85062-2989 Phoenix, United States
kliugastro@gmail.com

ABSTRACT

Background and study aims The Bethesda ERCP Skill Assessment Tool (BESAT) is a video-based assessment tool of technical endoscopic retrograde cholangiopancreatography (ERCP) skill with previously established validity evidence. We aimed to assess the discriminative validity of the BESAT in differentiating ERCP skill levels.

Methods Twelve experienced ERCP practitioners from tertiary academic centers were asked to blindly rate 43 ERCP videos using the BESAT. ERCP videos consisted of native biliary cannulation and sphincterotomy and were recorded from 10 unique endoscopists of various ERCP experience (from advanced endoscopy fellow to > 10 years of ERCP experience). Inter-rater reliability, discriminative validity, and internal structure validity were subsequently assessed.

Results The BESAT was found to reliably differentiate between endoscopists of varying levels of ERCP experience with experienced ERCPists scoring higher than novice ERCPists in 11 of 13 (85%) instrument items. Inter-rater reliability for BESAT items ranged from good to excellent (intraclass correlation range: 0.86 to 0.93). Internal structure validity was assessed with item-total correlations ranging from 0.53 to 0.83.

Conclusions Study findings demonstrate that the BESAT, a video-based ERCP skill assessment tool, has high inter-rater reliability and has discriminative validity in differentiating novice from expert ERCP skill. Further investigations are needed to determine the role of video-based assessment in improving trainee learning curves and patient outcomes.

Introduction

Endoscopic retrograde cholangiopancreatography (ERCP) is essential in the management of pancreaticobiliary diseases but is technically challenging and conveys increased risks (unsuccessful cannulation, post-ERCP pancreatitis, perforation, and bleeding) compared with most other endoscopic procedures [1, 2]. The advent of less invasive diagnostic testing, including endoscopic ultrasound (EUS) and magnetic resonance cholangiopancreatography (MRCP), has shifted the role of ERCP toward a predominantly therapeutic procedure, requiring all ERCPists to become competent in sphincterotomy, stent placement, and tissue sampling. As such, achieving competence in performing ERCP necessitates learning additional technical and cognitive skills beyond standard endoscopic procedures.

Standardized measures of competence in ERCP training have not been established. Similar to the history of colonoscopy training, which had previously emphasized volume thresholds and cecal intubation rate as markers of overall competence, assessment of ERCP competence has largely been determined based upon volume thresholds and, on occasion, determining native papilla biliary cannulation rates [3]. Based largely upon expert opinion, prior guidance suggested that ERCP competency be assessed after 200 procedures are performed with a goal of 90% achievement of selective native papilla deep cannulation [4].

In recent years, a paradigm shift from volume thresholds to competency-based metrics has begun to emerge for ERCP training. A growing literature base has shown significant variation in the rate at which trainees learn and acquire ERCP skills [5, 6]. Relying solely on procedure thresholds as a surrogate for ERCP competency overlooks the significant variations that can exist among both trainees and educators [7]. In response to this, the ERCP and EUS Skills Assessment Tool (TEESAT) was developed [6, 8, 9]. The TEESAT is a validated tool that was developed to enable trainers to assess trainee competence in performing all core and advanced ERCP skills, including achieving a stable duodenoscope position, successful cannulation and sphincterotomy, and ability to perform "advanced" cannulation techniques. While this work clearly confirmed the variable rates at which trainees achieve competence in performing ERCP, the TEESAT provides limited feedback to learners to foster skills improvement.

Video-based assessment has been demonstrated to be an effective and scalable tool for improving technical performance outcomes in surgical training [10, 11, 12, 13, 14]. Video-based assessment tools have also been developed and validated for colonoscopy and polypectomy [15, 16, 17]. Given the rise in readily available video capture technology, video-based skills assessment for complex endoscopic procedures has the potential to facilitate ERCP performance improvement through feedback provision and may help to overcome some potential challenges associated with direct observation assessment tools, including rater bias, the mental load associated with assessments in the clinical setting, and access to raters in smaller centers and/or low-resource settings [17]. However, a significant gap exists in development and implementation of video-based assessment measures. Recently, the Bethesda ERCP Skill Assessment Tool (BESAT), a novel tool for video-based assessment of ERCP skill, was developed with establishment of evidence for content validity, response process, and internal structure validity, namely reliability [15]. However, further validity evidence is required to substantiate use of the BESAT for the purpose of formative assessment to support video-based coaching and feedback. Thus, we aimed to evaluate additional validity evidence of BESAT as a formative video-based assessment tool using a recognized contemporary validity framework [18, 19]. The following sources of validity were appraised: 1) internal structure validity, that is, associations between BESAT items and how they relate to overall competence ratings; and 2) relations with other variables (i. e., discriminative validity), that is, the association of BESAT scores with experience level (expert/novice).

Methods

Previous development and validation of the BESAT

The BESAT was initially developed by nine principal investigators from the Stent Versus Indomethacin (SVI) trial with extensive ERCP experience [15]. Group members reviewed eight ERCP videos that included native papilla cannulation and sphincterotomy to deconstruct the procedure into its core

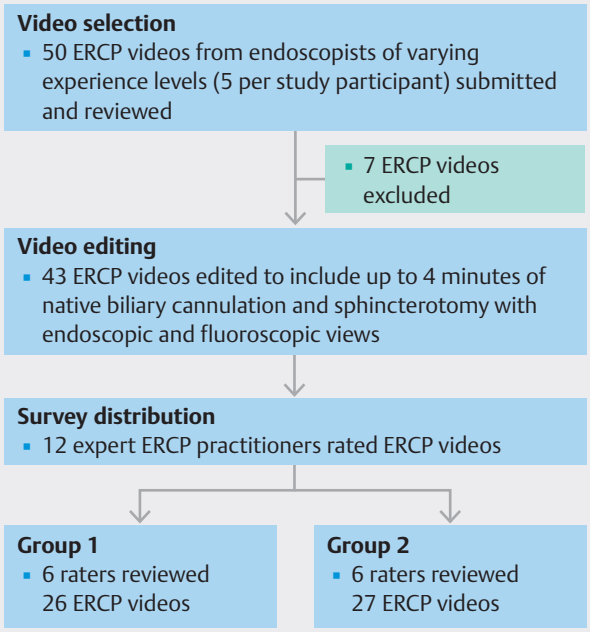
technical elements. A modified Delphi process and validation exercises were then used to iteratively refine the measure. The final BESAT tool is composed of six technical elements and 11 sub-elements of ERCP skill [15]. For example, the technical element of "alignment and maintenance of positioning" is composed of the two sub-elements "duodenoscope stability" and "papillary alignment."

Study design

This was a multicenter retrospective study. ERCP videos from endoscopists at three tertiary academic referral centers with high ERCP volumes that participated in the SVI trial were collected and analyzed. ERCPs in the selected videos were performed either by "novice" advanced endoscopy trainees with less than 4 months of ERCP experience or "experienced: advanced endoscopists with ≥ 12 months of supervised and/or independent ERCP experience. Study participants included two advanced endoscopy trainees (< 4 months of training), one graduating advanced endoscopy trainee (12 months of training), three junior advanced endoscopists (< 10 years of ERCP experience), and four senior advanced endoscopists (>10 years of ERCP experience).

All included videos involved native papilla biliary cannulation and sphincterotomy. ERCPs selected for analysis only included ERCPs with specific indications for biliary cannulation (i. e. choledocholithiasis, biliary anastomotic stricture after transplant, etc.). For the three ERCP trainees in the study (two novice and one experienced), only footage of the trainees independently attempting cannulation and sphincterotomy was included. However, the trainers may have given verbal feedback during the procedure to aid the trainee. For the seven independent endoscopist participants, only procedures that did not include a trainee were included. Analyzed videos included endoscopic and fluoroscopic footage from time of duodenoscope insertion to scope withdrawal. Initially, 50 ERCP videos (5 from each endoscopist) were selected and reviewed. Videos were excluded from the study for intentional pancreatic duct cannulation or lack of fluoroscopic images. Eligible videos were then de-identified and edited to include up to 4 minutes of each biliary cannulation and sphincterotomy. For prolonged biliary cannulation requiring > 4 minutes, cannulation footage was edited to highlight active cannulation efforts, with exclusion of periods of relative endoscopic and fluoroscopic inactivity (i. e., when the sphincterotome was not within the duodenum).

ERCP videos were then distributed electronically in survey format to 12 experienced ERCP practitioners in the United States for assessment using the BESAT. These raters worked at different institutions than the study participants, practiced at tertiary academic centers, and had more than 5 years of practice experience with more than 200 ERCPs performed annually. The 12 raters were split into two groups of six, each of which received two to three blinded ERCP videos per endoscopist from varying ERCP experience levels for assessment (▶ **Fig. 1**).



▶ **Fig. 1** Study design.

Study outcomes

The following outcomes were studied in accordance with Messick's unified validity framework, the current standard for evaluating validity evidence for performance assessments [19, 20]: 1) internal structure: inter-rater reliability, internal consistency, item-item and item-total correlations were examined; and 2) relations with other variables: associations of scores with level experience were evaluated (i. e., discriminative validity).

Statistical analysis

Data were exported from Qualtrics into SPSS v.27 for analyses. A BESAT Total Composite Score was calculated by adding all 12 items on the tool with exclusion of the "global assessment" score. Mean (SD) scores were computed for each technical element and sub-element on the BESAT. Tests for normal distribution (skewness and kurtosis +/- 2.0) were performed on all variables to determine the need for non-parametric tests. Descriptive statistics for the sample are presented as mean (SD) for continuous variables and percentage (frequency) for categorical; for non-normally distributed data median is reported with interquartile range.

Internal structure validity evidence

Internal structure validity of the BESAT was assessed using inter-item correlations with a goal range of 0.3 to 0.9 item-total correlations. Internal consistency was measured using Cronbach's alpha to indicate how well the BESAT items measures the intended overarching construct (i. e., competence in performing ERCP). Internal. Inter-rater reliability, or the level of agreement achieved by independent raters assessing the same performance, was also evaluated using intraclass coefficients.

Values below 0.5 indicated poor reliability, 0.5 and 0.75 moderate reliability, 0.75 and 0.9 good reliability, and above 0.9 indicated excellent reliability [21].

Relations with other variables validity evidence

To evaluate discriminative validity, the BESAT Total Composite score, in addition to the mean score across each BESAT element and sub-element was compared between the "novice" and "experienced" groups using independent samples *t*-tests. ERCP skill differences were measured with one-way ANOVA and Tukey post-hoc test. *P* was set to ≤.01 to control for Type 1 error due to multiple comparisons.

Results

Of the 50 ERCP videos reviewed, seven were excluded (4 for intentional pancreatic duct cannulation and 3 for lack of fluoroscopic images). BESAT ratings were completed for 43 ERCP videos from two novice ERCP practitioners (9 videos) and eight experienced ERCP practitioners (34 videos). The experienced ERCP practitioners had a mean of 7.9 ± 0.49 years of ERCP experience compared to the novice group with a mean 4 months of ERCP experience.

Internal structure validity evidence

The inter-rater reliability for each of the BESAT items ranged from good to excellent (intraclass correlation coefficient range: 0.86–0.93). Internal consistency of the BESAT tool was 0.95. Inter-item and item-total correlations are detailed in ▶ **Table 1**. The item-total correlations ranged from 0.53 to 0.83. Inter-item correlations between the "global assessment" and: 1) cannulation efficiency; 2) positioning and trajectory of catheter/wire; and 3) procedural judgment were highest at 0.84, 0.82 and 0.85, respectively.

Relations with other variables validity evidence: novice vs experienced performance

Total composite score, global assessment, and procedural judgment

The experienced ERCP group scored significantly higher compared to the novice ERCP group in the Total Composite Score (48.46 ± 7.17, vs 44.1 ± 8.1, $P < 0.05$), which was the sum of all BESAT technical skills. The experienced ERCP group also scored significantly higher in global assessment (3.6 ± 0.7 vs 2.86 ± 0.8, $P = 0.003$), which was the overall subjective assessment of endoscopist skill level. Similarly, the experienced ERCP group scored significantly higher in procedural judgment (3.78 ± 0.65 vs 3.29 ± 0.76, $P = 0.003$) compared to the novice ERCP group.

Alignment and maintenance of positioning

In the technical skill category of alignment and maintenance of positioning, the experienced ERCP group scored significantly higher in total mean score compared to the novice ERCP group (8.18 ± 1.28 vs 7.32 ± 1.22, $P = 0.006$). The experienced ERCP group also demonstrated significantly higher total mean scores compared to the novice ERCP group in sub-elements of duodenoscope stability (4.31 ± 0.61 vs 3.94 ± 0.75, $P = 0.017$) and papillary alignment (3.86 ± 0.77 vs 3.37 ± 0.76, $P = 0.01$; ▶ **Table 2**, ▶ **Fig. 2**).

Cannulation

In the technical skill category of cannulation, no significant difference was found in total mean score between experienced and novice ERCP groups (10.63 ± 2.16 vs 9.86 ± 1.81, $P = 0.13$). The experienced ERCP group scored significantly higher compared to the novice ERCP group in sub-elements of efficiency (3.63 ± 0.71 vs 2.95 ± 0.75, $P < 0.001$) and positioning and trajectory of catheter/wire (3.57 ± 0.84 vs 3.09 ± 0.64, $P = 0.006$). In contrast, the novice ERCP group scored significantly higher compared to the experienced ERCP group in the sub-element of gentleness of manipulation (3.82 ± 0.63 vs 3.43 ± 0.89, $P = 0.02$).

Sphincterotomy

In the technical skill category of sphincterotomy, the experienced ERCP group scored significantly higher in total mean score compared to the novice ERCP group (16.56 ± 2.62 vs 14.42 ± 2.46, $P = 0.001$). The experienced ERCP group scored significantly higher in all sphincterotomy sub-elements: control (4.15 ± 0.7 vs 3.69 ± 0.69, $P = 0.006$), trajectory (4.2 ± 0.71 vs 3.78 ± 0.77, $P = 0.016$), avoidance of excess diathermy injury/charring (3.98 ± 0.85 vs 3.45 ± 0.92, $p = 0.01$), and adequacy of size for indication (4.32 ± 0.78 vs 3.5 ± 0.73, $P < 0.001$).

Wire manipulation

In the technical skill category of wire manipulation, no significant difference was found in total mean score between the experienced and novice ERCP groups (8.24 ± 1.42 vs 7.71 ± 1.83, $P = 0.16$). The experienced ERCP group scored significantly higher compared to the novice ERCP group in the sub-element of stable and appropriate wire positioning (4.33 ± 0.66 vs 3.84 ± 0.93, $P = 0.007$). There was no significant difference between experienced and novice ERCP groups for the sub-element of wire advancement (3.86 ± 0.94 vs 3.79 ± 0.93, $P = 0.76$).

Secondary analyses

An analysis was performed comparing all trainees (2 trainees with < 4 months of ERCP experience and one graduating trainee with 12 months of ERCP experience) to independent ERCP practitioners. In this analysis, there was no significant difference in scores between the independent ERCP group and trainee ERCP group in all technical skill elements and sub-elements except for the sub-element of gentleness of manipulation (under the skill domain of cannulation), for which trainees scored significantly higher than the independent ERCPists (3.9 ± 0.63 vs 3.3 ± 0.89, $P < 0.001$).

A second analysis was performed comparing junior advanced endoscopists (< 10 years' experience) with senior advanced endoscopists (> 10 years' experience). There was no significant difference in scores between junior advanced endoscopists and senior advanced endoscopists in all technical skill elements and sub-elements.

▶ **Table 1** Inter-item and item-total correlations for BESAT.

	DS	PA	E	GP	PTC	C	T	AEDI	ASI	WA	SAWP	PJ	GA	TBS
Alignment and maintenance of positioning														
Duodenoscope stability	–													
Papillary alignment	.624	–												
Cannulation														
Efficiency	.618	.711	–											
Gentleness of manipulation	.402	.463	.644	–										
Positioning and trajectory of catheter/wire	.538	.782	.745	.674	–									
Sphincterotomy														
Control	.473	.401	.545	.437	.425	–								
Trajectory	.435	.397	.510	.430	.475	.742	–							
Avoidance of excess diathermy injury	.327	.323	.402	.291	.449	.618	.681	–						
Adequacy of size for indication	.466	.471	.660	.355	.575	.682	.645	.637	–					
Wire manipulation and positioning														
Wire advancement	.312*	.429	.449	.492	.577	.529	.555	.473	.399	–				
Stable and appropriate wire position	.382	.280*	.315*	.323	.361	.598	.573	.460	.407	.750	–			
Procedural judgment	.610	.709	.780	.629	.763	.578	.610	.493	.596	.655	.515	–		
Global assessment	.572	.735	.840	.611	.815	.552	.580	.483	.662	.550	.390	.849	–	
Total BESAT score	.593	.618	.639	.532	.578	.575	.727	.606	.742	.726	.696	.827	.744	–

BESAT, Bethesda ERCP Skill Assessment Tool; DS, duodenoscope stability; PA, papillary alignment; E, efficiency; GM, gentleness of manipulation; PTC, positioning and trajectory of catheter/wire; C, control; T, trajectory; AEDI, avoidance of excess diathermy injury; ASI, adequacy of size for indication; WA, wire advancement; SAWP, stable and appropriate wire position; PJ, procedural judgement; GA, global assessment; TBS, total BESAT score.
*$P < .05$; all other $P < .0$.

Discussion

The BESAT, a video-based tool for assessing competence in performing ERCP, was recently developed and shown to have evidence of content validity, response process, and internal structure validity, namely reliability [22]. Specifically, the BESAT was found to be a largely consistent tool wherein most of the variance was thought to be related to the performance of the endoscopist. Our study accrued validity evidence supporting relations with other variables by demonstrating the ability of the instrument to discriminate between the performance of "novice" and "experienced" endoscopists. As expected, the experienced ERCP group scored significantly higher than the novice group on 11 of 13 (85%) BESAT items. Thus, raters could watch approximately 8 minutes of a videotaped ERCP procedure and assess whether the endoscopist was more likely a novice or experienced practitioner.

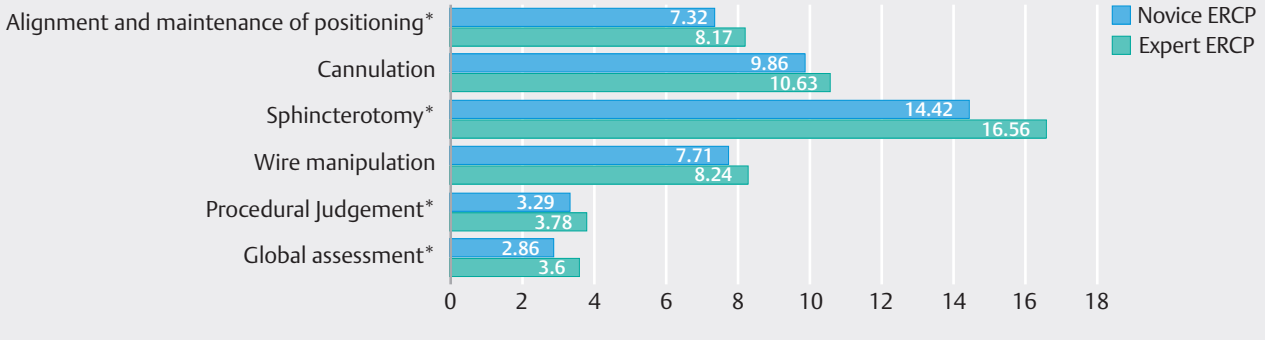
This study also established additional internal structure validity evidence supporting the BESAT. Internal consistency was excellent, providing support that the BESAT items provide a cohesive measure of competence in performing ERCP. In addition, as expected, the Total Composite BESAT score was highly correlated with overall global assessment of performance. Items pertaining to procedural judgment, trajectory, adequacy of size for endication and wire advancement correlated most strongly with Total Composite BESAT score. Additionally, inter-rater reliability was good to excellent (0.86–0.93), and considered acceptably high (> 0.75) for formative low-stakes assessment with minimal rater training [21,23].

TEESAT, an existing competency assessment tool that is rated based on direct observation, has strong validity evidence

▶ **Table 2** Mean BESAT technical skill element and sub-element scores and global assessment scores for novice and experienced ERCP practitioners.

Skill assessment*	Novice ERCPists (n = 2)	Experienced ERCPists (n = 8)	*P* value†
Alignment and maintenance of positioning (total of 2 sub-elements)	7.32 (1.22)	8.17 (1.28)	.006
Duodenoscope stability	3.94 (0.75)	4.31 (0.61)	.017
Papillary alignment	3.37 (0.76)	3.86 (0.77)	.010
Cannulation (total of 3 sub-elements)	9.86 (1.81)	10.63 (2.16)	.130
Efficiency	2.95 (0.75)	3.63 (0.71)	< .001
Gentleness of manipulation	3.82 (0.63)	3.43 (0.89)	.022
Positioning and trajectory of catheter/wire	3.09 (0.64)	3.57 (0.84)	.006
Sphincterotomy (total of 4 sub-elements)	14.42 (2.46)	16.56 (2.62)	.001
Control	3.69 (0.69)	4.15 (0.70)	.006
Trajectory	3.78 (0.77)	4.20 (0.71)	.016
Avoidance of excess diathermy injury/charring	3.45 (0.92)	3.98 (0.85)	.011
Adequacy of size for indication	3.50 (0.73)	4.23 (0.78)	< .001
Wire manipulation (total of 2 sub-elements)	7.71 (1.83)	8.24 (1.42)	.157
Wire advancement, including roadmap	3.79 (0.93)	3.86 (0.94)	.758
Stable and appropriate wire position	3.84 (0.93)	4.33 (0.66)	.007
Procedural judgment	3.29 (0.76)	3.78 (0.65)	.003
Composite score (total of all above)	44.09 (8.12)	48.46 (7.17)	.021
Global assessment	2.86 (0.80)	3.60 (0.70)	.001

BESAT, Bethesda ERCP Skill Assessment Tool; ERCP, endoscopic retrograde cholangiopancreatography.
*Individual skill assessment metrics were rated from 1 (worst) to 5 (best).
†P was set to .01 to control for Type 1 error due to multiple comparisons.



▶ **Fig. 2** Comparison of mean BESAT Technical Skill Element Scores between novice and experienced ERCP practitioners. *$P < 0.05$.

and has been recommended by the American Society for Gastrointestinal Endoscopy for assessment of technical, cognitive, and non-technical ERCP skills during training [5, 6, 7]. The BESAT is a video-based ERCP skills assessment tool that complements the TEESAT by obviating the need for direct observation and providing more specific feedback. By deconstructing ERCP into individual procedural sub-elements, the BESAT can provide unbiased and progressive feedback to foster skills development. For example, while the TEESAT can confirm that the trainee has not reliably achieved competence in cannulation, the BESAT can determine which individual skills that contribute to successful cannulation are not being achieved (e. g., not correctly positioning the papilla). The mental workload and attention capacity required by a rater to assess an encounter can be significant [24, 25]. In real-time procedure observations, raters are expected to process multiple streams of information and

transform these into multidimensional ratings, which can often to lead to biases in assessment [24]. Video-based assessment can help facilitate unbiased review of video after procedure completion without competing time constraints and environmental distractions.

Interestingly, the novice group scored higher than the expert group in Gentleness of Manipulation (under cannulation), which may reflect the difficulty trainees face in learning to appropriately manipulate the duodenoscope and sphincterotome before engaging the papilla (i. e., unable to adequately position the papilla to attempt cannulation) or the trepidation novice trainees might have in engaging the papilla. No difference was found between the novice and expert groups for wire advancement. These results may be due to difficulties in capturing subtleties in wire manipulation, which has a tactile component and cannot be appreciated purely based on fluoroscopic views. Furthermore, more videos of complex cases may be required to better assess the skill of wire advancement.

Several limitations of our study must be recognized. On subanalysis, the BESAT was not able to differentiate between a graduating trainee and attending ERCP practitioners. As there was only a single graduating trainee included in video analysis, this may be related to inadequate sample size rather than a limitation of the BESAT. Alternatively, this may reflect that these were truncated videos and that the first few minutes of cannulation attempts cannot adequately differentiate skill between experienced ERCPists. Specific indications for ERCP may also be more likely to be technically difficult (i. e. biliary cannulation for an obstructing pancreas head with duodenal distortion of the papilla) and future subgroup analyses based on indications for ERCP will be important. In addition, all trainee and attending videos were obtained from tertiary academic centers enrolled in the SVI trial and may not represent the variability of ERCP in general clinical practice [26]. Future studies incorporating a greater number of trainees and attendings would be helpful in assessing for these differences. To improve the feasibility of video review, we also opted to use edited videos for assessment in our study. Although edited videos for surgical assessment have previously been associated with worse inter-rater reliability, inter-rater reliability of the BESAT has consistently been found to be high [27, 28]. Lastly, important factors related to ERCP training including intraprocedure verbal feedback and coaching by supervising ERCPists as well as body and hand positioning were not able to be included in endoscopic and fluoroscopic video feedback. Future studies assessing the impact of external video feedback in addition to endoscopic and fluoroscopic video feedback may be helpful.

Conclusions

ERCP is a technically challenging procedure that carries significant risks and assessment tools should strive to support performance improvement, and ultimately, patient care. Our study provides evidence of internal structure and relations with other variables validity for the BESAT, a video-based ERCP skills assessment tool. Video-based assessment of ERCP and its implications for ERCP learning and training are still early in their development. These findings, combined with existing published evidence, augment the existing data in support of the BESAT as a formative assessment tool [22]. They also help to lay the foundation for future research to develop competency benchmarks and examine the impact of structured feedback based on BESAT on trainee learning curves and patient outcomes.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Vandervoort J, Soetikno RM, Tham TC et al. Risk factors for complications after performance of ERCP. Gastrointest Endosc 2002; 56: 652–656 doi:10.1067/mge.2002.129086

[2] Cotton PB, Garrow DA, Gallagher J et al. Risk factors for complications after ERCP: a multivariate analysis of 11,497 procedures over 12 years. Gastrointest Endosc 2009; 70: 80–88 doi:10.1016/j.gie.2008.10.039

[3] Shahidi N, Ou G, Telford J et al. Establishing the learning curve for achieving competency in performing colonoscopy: a systematic review. Gastrointest Endosc 2014; 80: 410–416 doi:10.1016/j.gie.2014.04.056

[4] ASGE Standards of Practice Committee. Faulx AL, Lightdale JR et al. Guidelines for privileging, credentialing, and proctoring to perform GI endoscopy. Gastrointest Endosc 2017; 85: 273–281

[5] 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

[6] Wani S, Hall M, Wang AY et al. Variation in learning curves and competence for ERCP among advanced endoscopy trainees by using cumulative sum analysis. Gastrointest Endosc 2016; 83: 711–719

[7] Wani S, Keswani RN, Petersen B et al. Training in EUS and ERCP: standardizing methods to assess competence. Gastrointest Endosc 2018; 87: 1371–1382 doi:10.1016/j.gie.2018.02.009

[8] Wani S, Cote 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

[9] 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 e1312 doi:10.1016/j.cgh.2014.11.008

[10] McQueen S, McKinnon V, VanderBeek L et al. Video-Based Assessment in Surgical Education: A Scoping Review. J Surg Educ 2019; 76: 1645–1654 doi:10.1016/j.jsurg.2019.05.013

[11] Augestad KM, Butt K, Ignjatovic D et al. Video-based coaching in surgical education: a systematic review and meta-analysis. Surg Endosc 2020; 34: 521–535 doi:10.1007/s00464-019-07265-0

[12] Netter A, Schmitt A, Agostini A et al. Video-based self-assessment enhances laparoscopic skills on a virtual reality simulator: a randomized controlled trial. Surg Endosc 2021; 35: 6679–6686 doi:10.1007/s00464-020-08170-7

[13] Mota P, Carvalho N, Carvalho-Dias E et al. Video-based surgical learning: improving trainee education and preparation for surgery. J Surg Educ 2018; 75: 828–835 doi:10.1016/j.jsurg.2017.09.027

[14] Rindos NB, Wroble-Biglan M, Ecker A et al. Impact of video coaching on gynecologic resident laparoscopic suturing: a randomized controlled trial. J Minim Invasive Gynecol 2017; 24: 426–431 doi:10.1016/j.jmig.2016.12.020

[15] Elmunzer BJ, Walsh CM, Guiton G et al. Development and initial validation of an instrument for video-based assessment of technical skill in ERCP. Gastrointest Endosc 2021; 93: 914–923 doi:10.1016/j.gie.2020.07.055

[16] Scaffidi MA, Grover SC, Carnahan H et al. A prospective comparison of live and video-based assessments of colonoscopy performance. Gastrointest Endosc 2018; 87: 766–775 doi:10.1016/j.gie.2017.08.020

[17] Jeyalingam T, Walsh CM. Video-based assessments: a promising step in improving polypectomy competency. Gastrointest Endosc 2019; 89: 1231–1233 doi:10.1016/j.gie.2019.04.203

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

[19] Messick S. Educational measurement 1989; 3rd ed, p. 13–103.

[20] Kane M. An argument-based approach to validity. Psychol Bull 1992; 112: 527–535

[21] Downing S. Reliability: on the reproducability of assessment data. Med Educ 2004: doi:10.1111/j.1365-2929.2004.01932.x

[22] Elmunzer BJ, Walsh CM, Guiton G et al. Development and initial validation of an instrument for video-based assessment of technical skill in ERCP. Gastrointestinal endoscopy 2021; 93: 914–923 doi:10.1016/j.gie.2020.07.055

[23] Watkins M. Foundations of clinical research: applications to practice. Pearson Education Inc.. 2009

[24] Tavares W, Eva KW. Exploring the impact of mental workload on rater-based assessments. Adv Health Sci Educ Theory Pract 2013; 18: 291–303 doi:10.1007/s10459-012-9370-3

[25] Tavares W, Ginsburg S, Eva KW. Selecting and simplifying: rater performance and behavior when considering multiple competencies. Teach Learn Med 2016; 28: 41–51 doi:10.1080/10401334.2015.1107489

[26] Elmunzer BJ, Serrano J, Chak A et al. Rectal indomethacin alone versus indomethacin and prophylactic pancreatic stent placement for preventing pancreatitis after ERCP: study protocol for a randomized controlled trial. Trials 2016; 17: 120 doi:10.1186/s13063-020-04458-0

[27] Scott DJ, Rege RV, Bergen PC et al. Measuring operative performance after laparoscopic skills training: edited videotape versus direct observation. J Laparoendosc Adv Surg Tech A 2000; 10: 183–190 doi:10.1089/109264200421559

[28] Sawyer JM, Anton NE, Korndorffer JR et al. Time crunch: increasing the efficiency of assessment of technical surgical skill via brief video clips. Surgery 2018; 163: 933–937 doi:10.1016/j.surg.2017.11.011