Accepted Manuscript online: 2022-04-01 Article published online: 2022-06-10

Assessment of esophagogastroduodenoscopy skills on simulators before real-life performance

OPEN ACCESS

Authors
Anders Bo Nielsen[1,2,3], Finn Møller Pedersen[2,3], Christian B. Laursen[4,5], Lars Konge[6], Stig Laursen[2,3]

Institutions
1 Odense University Hospital, SimC – Simulation Center, Odense, Denmark
2 Odense University Hospital, Department of Medical Gastroenterology, Odense, Denmark
3 University of Southern Denmark, Department of Clinical Research, Odense, Denmark
4 Odense University Hospital, Department of Respiratory Medicine, Odense, Denmark
5 University of Southern Denmark, Respiratory Research Unit, Odense, Denmark
6 Capital Region of Denmark – Copenhagen Academy for Medical Education and Simulation, Copenhagen, Denmark

submitted 20.10.2021
accepted after revision 30.3.2022
published online 1.4.2022

Bibliography
Endosc Int Open 2022; 10: E815–E823
DOI 10.1055/a-1814-9747
ISSN 2364-3722

Georg Thieme Verlag KG, Rüdigerstraße 14,
70469 Stuttgart, Germany

Corresponding author
Anders Bo Nielsen, Odense University Hospital – SimC – Simulation Center, J. B. Winsløw Vej 4, Odense 5000, Denmark
Phone: +4531378203
anders.bo.nielsen@rsyd.dk

ABSTRACT

Background and study aims Operator competency is essential for esophagogastroduodenoscopy (EGD) quality, which makes appropriate training with a final test important. The aims of this study were to develop a test for assessing skills in performing EGD, gather validity evidence for the test, and establish a credible pass/fail score.

Methods An expert panel developed a practical test using the Simbionix GI Mentor II simulator (3D Systems) and an EGD phantom (OGI 4, CLA Medical) with a diagnostic (DP) and a technical skills part (TSP) for a prospective validation study. During the test a supervisor measured: 1) total time; 2) degree of mucosal visualization; and 3) landmarks and pathology identification. The contrasting groups standard setting method was used to establish a pass/fail score.

Results We included 15 novices (N), 10 intermediates (I), and 10 experienced endoscopists (E). The internal structure was high with a Cronbach's alpha of 0.76 for TSP time consumption and 0.74 for the identification of landmarks.
Mean total times, in minutes, for the DP were N 15.7, I 11.3, and E 7.0, and for TSP., they were N 7.9, I 8.9, and E 2.9. The total numbers of identified landmarks were N 26, I 41, and E 48. Mean visualization percentages were N 80, I 71, and E 71. A pass/fail standard was established requiring identification of all landmarks and performance of the TSP in <5 minutes. All experienced endoscopists passed, while none of the endoscopists in the other categories did.

Conclusions We established a test that can distinguish between participants with different competencies. This enables an objective and evidence-based approach to assessment of competencies in EGD.

Introduction

Esophagogastroduodenoscopy (EGD) is one of the cornerstones for evaluation of patients with symptoms originating from the upper gastrointestinal tract, including heartburn, dyspepsia, upper abdominal pain, dysphagia, weight loss, hematemesis, and melena. EGD enables high-quality assessment of the mucosa in the upper gastrointestinal tract and makes it possible to obtain biopsies, perform endoscopic treatment, and retrieve foreign bodies [1].

EGD training programs are often based on traditional apprenticeship learning and the "see one, do one, teach one" method, which is not keeping with the concept of putting patient safety first [2]. It would be beneficial for endoscopists to do simulation-based technical skills training on scope and tool handling in a risk-free environment to gain dexterity in scope movements and orientation between the three-dimensional gastrointestinal tract and the two-dimensional screen [3]. The latest Cochrane review on virtual reality (VR) simulation training in endoscopy emphasizes that simulation can provide better educational programs [4].

Simulation-based assessments for which evidence of validity in EGD has been systemically gathered are limited [5–7]. Competence assessment in these studies was mainly based on obtaining a predefined number of training hours or completion of a predefined number of cases, which is inferior to mastery learning involving a test with a pass/fail score [8, 9]. Essentials of mastery learning are to develop training programs of high-quality, with predefined learning objectives and clear assessment parameters to ensure standard levels of skills [9–11]. Trainees will then reach the same minimum level of skills before embarking on the clinical training program as a part of their learning curve [12, 13].

The aims of this study were to: 1) develop a test for assessing skills in performing EGD; 2) gather validity evidence for the test; and 3) establish a credible pass/fail score for ensuring a baseline of EGD skills prior to clinical training.

Methods

Setting

The study was designed as a prospective validation study and carried out at the regional center for technical simulation (SimC) at Odense University Hospital, Region of Southern Denmark.

Validity evidence

The principles and framework of Messick [14, 15] were used to gather validity evidence for a test with the five sources of evidence: content, response process, internal structure, relationship to other variables, and consequences of testing [16].

Simulation-based test

Two EGD experts (SBL + FMP), a professor in medical education and simulation (LK), and a simulation expert (ABN) evaluated the feasibility and clinical relevance of the diagnostic cases and scope-handling exercises of the Simbionix GI Mentor II (3D systems, California) VR simulator. A consensus was reached on a test including an introduction case with a healthy patient (Module 1, Case 1), a diagnostic case with a hiatal hernia and an esophageal diverticulum (Module 1, Case 2), and a case with a fundic tumor (Module 1, Case 8). Moreover, tool handling was tested using the EndoBubble Case 1 (popping 20 balloons in a pipe with the scope) repeated three times (▶ **Fig. 1a**, ▶ **Fig. 1b**, ▶ **Fig. 1c**).

Finally, the panel developed a test for the OGI CLA four phantom with a real-life gastroscope (Olympus Exera 2 CV-180 Video Endoscopy System) including: 1) a diagnostic EGD of the phantom and two tool handling exercises; 2) retrieval of a suture (5-cm Ethicon Mersilene CP-2 0) placed at the greater gastric curvature; and 3) retrieval of a plastic bead (5-mm blue bead) in the gastric antrum (▶ **Fig. 1d**, ▶ **Fig. 1e**, ▶ **Fig. 1f**). The retrieval forceps had mixed teeth (MicroTech Type Griffin: Long alligator jaw with 2:1 teeth).

A pilot test was carried out and all cases were completed in a satisfactory manner before the final test was decided upon. Three novices, one intermediate, and one experienced endoscopist were enrolled in the pilot study and they had one attempt each to complete the test. None of the results or participants from the pilot study were included in the final data collection.

The simulator software was not updated throughout the study.

Participants

Three groups with different levels of EGD experience were enrolled in the study. Novices were medical students with no EGD experience who had passed their anatomy exams. Intermediates were endoscopy-assisting nurses who never performed an EGD but had assisted with > 500 EGDs. Experienced endoscopists were medical doctors in gastroenterology or surgery who had performed > 500 EGDs.

Novices were enrolled at the University of Southern Denmark. Groups of intermediates and experienced were recruited from the Department of Gastroenterology or Department of Surgery at Odense University Hospital, Denmark. Prior experience with EGD simulation was an exclusion criterion.

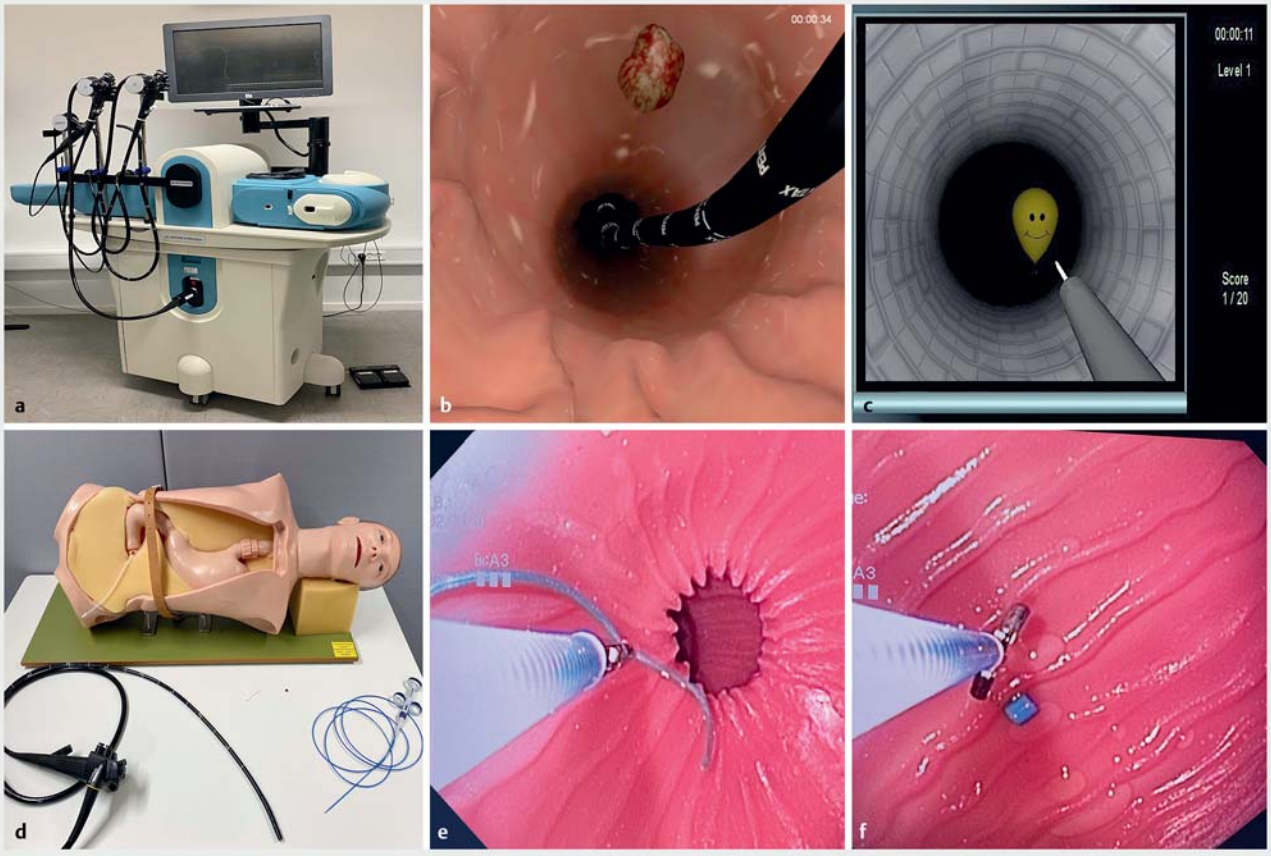
We aimed to include a minimum of 10 participants per group to fulfill the assumption of normally distributed data in medical educational research [17].

Completion of test and data collection

Validity evidence toward the response process was ensured by standardizing the testing situation.

Each participant received a unique identification number to preserve their anonymity and they received an identical introduction read aloud from an instruction paper about the aim, test program, simulators, and anatomy of the upper gastrointestinal tract.

To ensure familiarization with the simulator, the 5-minute introduction case was not included in the test program. Participants were asked to identify 15 landmarks (▶ **Table 1**) per diagnostic case and three pathologies in total. Simulator metrics were collected, including total time for the examination, measured from intubation to extraction of the scope, percentage of total mucosal surface examined, efficiency of screening (percentage of mucosal surface visualized in relation to time), and percentage of time spent with a clear view. In addition, the number of popped balloons (up to 20), and wall hits (hitting the wall of the virtual pipe with the endoscope) were collected from the EndoBubble cases. Time spent on the cases with the phantom was measured with a stopwatch. No evaluation was given to the participants while the test was running.



▶ **Fig. 1** **a** The Simbionix GI Mentor II virtual reality simulator. **b** Fundus tumor in the diagnostic part. **c** EndoBubble. **d** Phantom-based setup. **e** Suture retrieval. **f** Bead retrieval.

Test score

During the test, the participants were asked to mark the identified landmarks and pathologies. Identification was continuously evaluated by ABN and scored correct, not correct, or missing.

Each landmark and pathology recognition equaled one point and the cumulative maximum score was 48 points. Simulator-generated metrics were collected after each case by ABN. Time consumption was also measured by ABN.

Statistical analyses

Statistical analysis was carried out in SPSS statistics version 26 (IBM, New York, United States).

The null hypothesis was that the groups of novices and experienced endoscopists would perform equally well on each of the included parameters (landmark and pathology recognition, time spent on the diagnostic and tool handling part, mucosal visualization percent, efficiency, and clear view).

To gather validity evidence about the test, the internal consistency reliability of the test was calculated using Cronbach's alpha. Internal one-way analysis of variance (ANOVA) was used to compare the test results between the groups. Bonferroni was used as a correction for multiplicity post hoc, including three dependent variables (landmark and pathology recognition, and time spent on the diagnostic and tool handling part). An Alpha value of 0.05 was used and the familywise error rate was 60 %. Pearson's *r* was used to determine correlation. A pass/fail score was established with the contrasting groups standard setting method, using the intersection between the performances of the novices and the experienced endoscopists [15]. Consequences of the pass/fail score for each of the three groups were shown in a 3×2 contingency table and analyzed using Fisher's exact test.

Corrected two-tailed $P<0.05$ was considered statistically significant.

Results

Enrollment and data collection were performed from October to November 2019. A total of 35 participants took part in the test, including 15 medical students, 10 endoscopy-assisting nurses, and 10 experienced endoscopists who were registrars or specialists in gastroenterology or surgery (▶ **Table 2**).

Internal structure

The internal consistency reliability of landmark and pathology identification showed a Cronbach alpha of 0.74. The same internal consistency was calculated for the spent in the tool handling part, showing a Cronbach's alpha of 0.76.

▶ **Table 1** Test content, findings, landmarks, and points.

Test elements	Modality	Case number	Findings (points)	Landmarks for each case (points)	
Introduction	Simulator	1; module 1	Normal		
Diagnostics	Simulator	2; module 1	Esophageal diverticulum (1) Hiatal hernia (1)	Stomach Cardia (1) Fundus (1) Greater curvature (1) Lesser curvature (1) Anterior wall (1) Posterior wall (1) Antrum (1) Angular incisure (1)	Duodenal bulb Anterior wall (1) Posterior wall (1) Roof of the duodenal bulb (1) Floor of the duodenal bulb (1) Superior duodenal flexure (1) Descending part of duodenum (1) Esophagus (1)
	Simulator	8; module 1	Fundus tumor (1)		
	Phantom	EGD			
Tool handling	Simulator	1, EndoBubble	20 balloons		
	Phantom	Suture retrieval			
	Phantom	Bead retrieval			

▶ **Table 2** Participant demographics.

	Novices	Intermediates	Experienced
Total, n	15	10	10
Female, %	67	100	30
Mean age, years (range)	26 (21–30)	45 (35–53)	49 (32–75)
Mean number of performed EGD (range)	0	0	7,420 (500–18,000)
Mean number of assisted EGD (range)	0	15,150 (900–50,000)	0

EGD, esophagogastroduodenoscopy.

The intraclass correlation coefficient (ICC) for landmark and pathology recognition in the diagnostic part in a single case was 0.48, and in a single case of the tool-handling part, it was 0.51, but in total for all diagnostic cases it was 0.76 and for all cases in the tool handling part, it was 0.93. This shows a high level of consistency in participant performance and a low risk of getting the score by chance [18].

The correlation between the two parts of the test had a Pearson's *r* linear value of 0.49, indicating slightly low reliability between the two parts of the test.

The ICCs for the simulator-generated metrics for the two cases were the percentage of total mucosal surface examined, which was 0.45, efficiency of screening 0.20, and the percentage spent with clear view was 0.44. Correlations between the two VR cases indicated no reliability within the simulator-generated metrics assessed by Pearson's *r* linear value 0.14 compared to tool handling time and recognition of landmarks/pathology.

Relations to other variables

The results of the test are shown in ▶ **Table 3**. Mean total times (minutes) for the diagnostic part were N 15.7 (95 % CI: 13.9–17.4), I 11.3 (95 % CI: 10.3–12.3), and E 7.0 (95 % CI: 5.5–8.5), and on the technical skills part N 7.9 (95 % CI: 5.5–10.4), I 8.9, (95 % CI: 7.6–10.1), and E 2.9 (95 % CI: 2.3–3.5). The total numbers of diagnostic landmarks and pathology identification were N 26 (95 % CI: 21–31), I 41 (95 % CI: 36.6–45.4), and E 48 (95 % CI: 48–48). Visualization percentages in Case 1 were N 74 (95 % CI: 69.3–78.0), I 69 (95 % CI: 64.8–73.4), and E 65 (95 % CI: 61.3–67.9), and for Case 2 were N 86 (95 % CI: 83.3–89.0), I 73 (95 % CI: 64.3–81.3), and E 77 (95 % CI: 69.2–85.3).

One-way ANOVA showed statistical significance for three metrics (landmark and pathology recognition $P<0.001$, duration of diagnostic part $P<0.001$, and duration of the tool handling part $P<0.001$). Including Bonferroni correction, significant differences were shown for experienced and novices in landmark/pathology identification (48 vs 26 points; $P<0.001$) and total time spent on the tool-handling part, including the average time spent on the EndoBubble tasks and retrieval of the suture and plastic bead (2.9 vs 7.9 min; $P<0.001$).

No difference among the groups were demonstrated for other parameters, such as percentage of mucosal surface examination, efficiency of screening (by percentage), and percentage of time spent with a clear view.

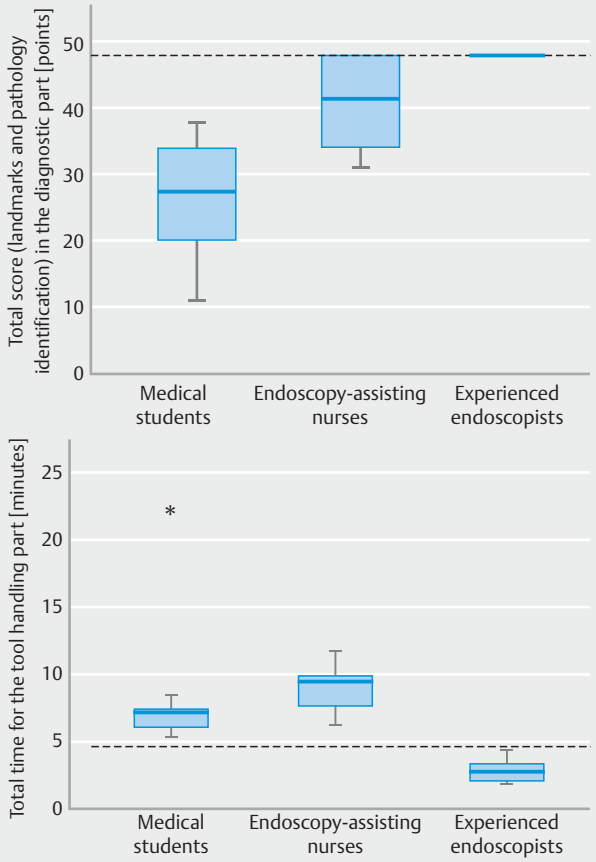
Performances of each of the groups on both the diagnostic and tool-handling part are illustrated in ▶ **Fig. 2**.

There was no statistically significant mean difference between endoscopy-assisting nurses and the experienced endoscopists or the endoscopy-assisting nurses and medical students, except in recognition of landmarks and pathology (20 points; $P=0.001$).

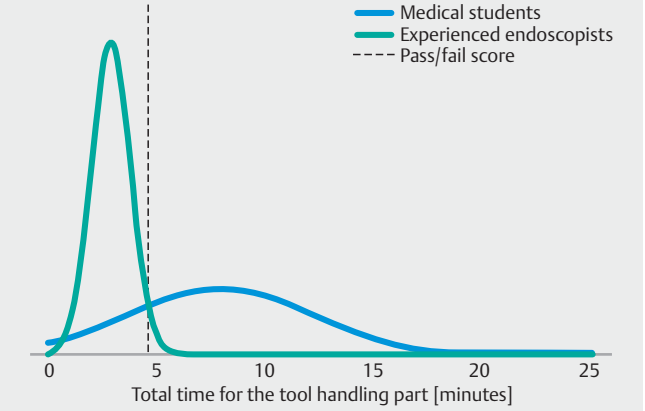
▶ **Table 3** Test performance among the three groups.

	Group (1)	N	Mean	SD	95 % CI for mean		Score		ANOVA	ANOVA[1]	Multiple comparisons[1]	
					Lower bound	Upper bound	Minimum	Maximum	Between groups (*P* value)		Groups	*P* value
Landmark and pathology recognition (points)	Novices	15	26	8.6	21.0	31.0	11	38	<0.001	<0.001	Novices vs Intermediates	<0.001
	Intermediates	10	41	6.1	36.6	45.4	31	48			Intermediates vs Experienced	0.051
	Experienced	10	48	0	48	48	48	48			Experienced vs Novices	<0.001
Time diagnostic part, (minutes)	Novices	15	15.7	3.1	13.9	17.4	9.1	20.6	<0.001	<0.001	Novices vs Intermediates	<0.001
	Intermediates	10	11.3	1.4	10.3	12.3	9.5	13.6			Intermediates vs Experienced	<0.001
	Experienced	10	7.0	2.2	5.5	8.5	4.0	12.1			Experienced vs Novices	<0.001
Time tool handling part, (minutes)	Novices	15	7.9	4.2	5.5	10.4	5.3	22.2	<0.001	<0.001	Novices vs Intermediates	1.000
	Intermediates	10	8.9	1.7	7.6	10.1	6.2	11.7			Intermediates vs Experienced	<0.001
	Experienced	10	2.9	0.9	2.3	3.5	1.8	4.4			Experienced vs Novices	<0.001
Visualization, (%) (simulator metric)	Novices	15	79.6	5.2	76.6	82.7	66.5	87.5	0.002	0.032	Novices vs Intermediates	0.007
	Intermediates	10	71.0	7.6	65.5	76.4	58.5	83.5			Intermediates vs Experienced	1.000
	Experienced	10	71.0	6.6	66.5	75.4	62.5	80.5			Experienced vs Novices	0.006
Efficiency (%)	Novices	15	76.4	7.1	72.4	80.5	60	85	0.178	1.000		
(simulator metric)	Intermediates	10	69.7	10	62.5	76.8	50.5	84				
	Experienced	10	72.8	9.3	66.5	79	60	89				
Clear View (%)	Novices	15	99.7	0.3	76.4	99.5	99	100	0.562	1.000		
(simulator metric)	Intermediates	10	99.5	0.5	99.1	9.9	98.5	100				
	Experienced	10	99.5	0.9	99.4	100	97	100				

CI, confidence interval; ANOVA, analysis of variance ; SD, standard deviation.
[1] Bonferroni corrected



▶ **Fig. 2** Box-and-whiskers plot illustrating the total point score for identification of landmarks and pathology in the diagnostic part and total time for the tool handling part for the three groups. Median, maximum, and minimum time/score are depicted. The dashed line shows the pass/fail score.



▶ **Fig. 3** Pass/fail score for total time use for the tool handling part illustrated by the dashed line using the contrasting groups method.

Consequences of testing

Using the contrasting groups standard setting method, a pass/fail score was established, based on the mean scores of the novices and experienced endoscopists. The pass/fail score consisted of two parts, both of which needed to be passed: 1) recognition of all landmarks (3 cases with 15 landmarks each) and three pathologies (1 point each) for 48 points in total; and 2) total maximum time for the two tool-handling tests of 4.6 minutes (▶ **Fig. 3**). None of the novices (i. e. no false positives) or intermediates passed the test while all experienced endoscopists passed (i. e. no false negatives). The theoretical false-positive rate was 0.5 % and theoretical false-negative rate was 0.0 % for landmark/pathology recognition, while the theoretical false-positive rate was 78.6 % and the theoretical false-negative rate was 97.3 % for time spent on the tool-handling part.

Discussion

We developed a practical, standardized simulation-based test with supporting validity evidence according to Messick's five sources of evidence. The test included both diagnostic and tool-handling part and demonstrated a reliable and valid approach for assessing EGD competency. The test discriminated between experience levels based on landmark/pathology recognition and time spent on the tool-handling part. To our knowledge, no other studies have gathered validity evidence for a simulation-based test to ensure basic competence in EGD.

According to the first of Messick's five sources, content, development of the test was provided by an experienced panel in EGD and simulation. The panel was asked to evaluate the content of the simulator and define the most relevant exercises. This approach has previously been used to obtain content consensus with success [19]. The risk of selection bias is an obvious risk when choosing an expert panel. We tried to overcome this potential risk by enrolling experts in various fields and from different departments (medical education, gastroenterology, surgery, simulation), but including a Delphi process would have increased the content validity [20].

To accommodate validity for the second source, the response process, all participants received the same written instructions and had the same introduction case with a time limit of 5 minutes training by the supervisor (ABN). During the tasks, no interaction between the participant and the supervisor was allowed to minimize the bias of an instructor-participant relationship together affecting the data. It would have been ideal to use simulator metrics for more objective assessment. However, the simulator was not capable of testing landmark recognition.

The third of Messick's sources, internal structure, involves meaningful interpretation and assessment of data and requires a Cronbach's alpha above 0.70 to obtain consistency and reproducibility of tests [21]. Our test had a Cronbach's alpha value between 0.74 and 0.76. with time spent on the diagnostic part being the most reliable measurement statistically, but the recommendation is that ≥ 7 minutes be spent on diagnostic EGD to achieve sufficient visualization [22, 23]. Therefore, time spent on the diagnostic part was not included as a requirement to be passed in the test. According to our findings, the correlation of simulator-generated metrics is weak, which is why they

cannot be used as parameters for assessing competency in EGD ability to perform reliable evaluation because the generated results would not be consistent [15].

According to relations to other variables, the fourth source, we established a test that can distinguish between different groups with various experience levels. A significant difference was shown in time spent on the cases and the ability to recognize landmarks and pathology between the experienced group and the novices. There was no significant difference between the intermediates and the other groups. The mean scores correlate with the experience level, which is in keeping with an increase in consistency and decrease in variance over time. This correlation fits with Fitts and Posner's theory about the three sequential stages of learning progress for motor skills [24]. The novices were in the cognitive trial-and-error phase (first stage). The second learning stage is 'the associative," in which participants such as the endoscopy-assisting nurses are competent with know-how about the procedure because they have assisted with multiple examinations, but lack any experience in handling the scope themselves. The third stage with autonomous skills corresponds to the experienced endoscopists' learning stage, including accurate and internalized performance. Compared to our results in the experienced group, their performance was homogenous in both time and landmark recognition [25, 26].

The discriminatory capability of simulator-generated metrics based on the Simbionix GI Mentor II is questionable. The Simbionix GI Mentor II simulator could not significantly discriminate among the groups based on integrated metrics including visualization of the mucosa, efficiency of screening percent, and percent of time spent with a clear view. We evaluated the metrics within the three groups and visualization of the mucosa was highest for the novices. However, the novices performed an unstructured evaluation according to international recommendations with almost no "red-out" in the simulator-generated metric of "clear view" [27]. The same tendency is seen in the predefined simulator metrics by Simbionix "efficiency of screening" (mucosal surface visualized per time), which was surprisingly high for all groups, even though most of the novices expressed confusion about their location in the gastrointestinal tract during the examination. An objective simulator metrics-based evaluation of competency is not possible because the simulator cannot discriminate correctly between levels of competence, and as a result, could not evaluate whether an operator had passed or failed the test [28].

The last of Messick's sources are the consequences of the test and the reliability of the test to set a mastery learning standard. We can rely on the content of the test because it discriminates among levels of competency [29]. A gold standard for determining scores and setting a guarantee for a clear-cut pass/fail score is not yet available to define those who are qualified to pass [15]. But it is well known that it is beneficial to test for competency and have a predefined score to be passed [30]. The contrasting groups standard setting method was used to calculate a pass/fail score as in other similar studies regarding technical skills acquisition [31].

At our hospital, we have previously used training programs for learning EGD that were based on obtaining predefined scores for simulator metrics such as mucosa visualization percentage. Our data indicate that use of mucosa visualization percentage as the main factor in evaluating skills in EGD may be fruitless, given the low discriminatory capability of the simulator. According to our results, the same challenge seems to apply to the other simulator-generated metrics. Therefore, the simulator is not reliable for making a proper decision about pass/failure of competency and an examiner should be used instead [5]. It is important to use mastery learning tests with gathered validity to ensure correct testing, keeping in mind the need to prioritize clinical relevance over statistical significance. Using a test based on metrics without validity evidence has potential to be dangerous to patients.

We aimed to develop and gather validity evidence for a test to assess competency in EGD. Our priority was to create a test based on metrics to avoid rater bias. This was not possible, given the results of the metrics and the discriminatory capability of the simulator. The participants were asked to mark 15 predefined landmarks in each of the three diagnostic cases. The markings were ticked off by the observer. On the basis of the results, we developed a test for marking diagnostic landmarks and of tool-handling skill.

Our test focused on the technical part of performing an EGD and not on the clinical setting with staff, patient care, and administrative work, which are also important parts of being a well-qualified examiner. Including these aspects may make the test more challenging and improve the mastery standard for learning EGD as a supplement to conventional clinical training [29]. Other studies have previously described simulation tasks as of too poor quality and with a lack of realism, including haptic and visual feedback, but in this study, it can be argued that on our test, the quality of the measurement of skill level discrimination for landmark/pathology recognition and tool handling was acceptable because all experts identified the landmarks/pathology and passed [5–7].

This study differs from other studies focusing on competency assessment based on simulation, given the focus on landmark and pathology recognition, time spent, and tool handling instead of completion of a given number of cases or training sessions [32–34]. This study focuses on reaching a specific level of competency in EGD using mastery learning. The same approach was used to develop and assess competency in other endoscopy procedures, such as bronchoscopy and colonoscopy [35, 36]. Our test assesses EGD competency exclusively in contrast to the SAGES (Society of American Gastrointestinal and Endoscopic Surgeons) Fundamentals of Endoscopic Surgery (FES) exam, which assesses gastrointestinal endoscopic skills including both EGD and colonoscopy [37]. Our test may be beneficial for educational programs that require learning and training in EGD and colonoscopy in different courses.

The strengths of our study include the use of Messick's framework and the setting of mastery learning standards. We aimed to develop a test using simulator-generated metrics for objective evaluation. These metrics were evaluated for clinical use and selected for the test by an expert panel within the field.

Moreover, we decided to evaluate landmark and pathology recognition by the observer because the simulator was not capable of these registrations. A weakness of this measurement method is that the observer was not blinded to participant experience level, but registrations of landmarks in the upper gastrointestinal tract are relatively simple and correctness of marking is easy to determine. Bias can be reduced, if using simulator metrics, but those are solely able to test technical skills [15]. A limitation is the use of endoscopy-assisting nurses as intermediates because they have limited scope-handling experience. In a future trial, it would be preferable to enroll a group of endoscopists with intermediate experience (e. g., 20–50 previous EGDs) to investigate performance of the test in real-life intermediate endoscopists. Another limitation is also the risk of lack of familiarity with the simulator. It might have been beneficial to give participants more than 5 minutes of rehearsal with the device. Similarly, sequential attempts at the test would have strengthened the investigation of this limitation and clarified the need for familiarization.

It is important to emphasize that this study can only be used to assess the technical skills for EGD in a simulation environment. Moreover, it is important to provide training on indications, contraindications, and clinical knowledge as well as non-technical skills separately. The clinical impact of passing this test needs to be evaluated in another clinical study.

Conclusions

In conclusion, we developed a simulation-based test for assessment of competence in EGD and established validity evidence for the test. The test can discriminate between groups with different experience levels with acceptable reliability. The established pass/fail standard resulted in no false positives or false negatives. This standardized test could be a prerequisite in a structured mastery learning training program.

Competing interests

The authors declare that they have no conflict of interest.

Clinical trial

ClinicalTrials.gov
NCT04150237
TRIAL REGISTRATION: Prospective study at https://www.clinicaltrials.gov/

References

[1] Beg S, Ragunath K, Wyman A et al. Quality standards in upper gastrointestinal endoscopy: a position statement of the British Society of Gastroenterology (BSG) and Association of Upper Gastrointestinal Surgeons of Great Britain and Ireland (AUGIS). Gut 2017; 66: 1886–1899

[2] Kotsis SV, Chung KC. Application of the “see one, do one, teach one” concept in surgical training. Plast Reconstr Surg 2013; 131: 1194–1201doi:10.1097/PRS.0b013e318287a0b3

[3] Ekkelenkamp VE, Koch AD, de Man RA et al. Training and competence assessment in GI endoscopy: a systematic review. Gut 2016; 65: 607–615

[4] Khan R, Plahouras J, Johnston BC et al. Virtual reality simulation training for health professions trainees in gastrointestinal endoscopy. Cochrane Database Syst Rev 2018; 8: Cd008237

[5] Sedlack RE. Validation of computer simulation training for esophagogastroduodenoscopy: Pilot study. J Gastroenterol Hepatol 2007; 22: 1214–1219

[6] McConnell RA, Kim S, Ahmad NA et al. Poor discriminatory function for endoscopic skills on a computer-based simulator. Gastrointest Endosc 2012; 76: 993–1002

[7] Ferlitsch A, Schoefl R, Puespoek A et al. Effect of virtual endoscopy simulator training on performance of upper gastrointestinal endoscopy in patients: a randomized controlled trial. Endoscopy 2010; 42: 1049–1056

[8] Lineberry M, Soo ParkY, Cook DA et al. Making the case for mastery learning assessments: key issues in validation and justification. Acad Med 2015; 90: 1445–1450

[9] Cook DA, Brydges R, Zendejas B et al. Mastery learning for health professionals using technology-enhanced simulation: a systematic review and meta-analysis. Acad Med 2013; 88: 1178–1186

[10] Cook DA, Hatala R, Brydges R et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 2011; 306: 978–988

[11] McGaghie WC. Mastery learning: it is time for medical education to join the 21st century. Acad Med 2015; 90: 1438–1441

[12] Cohen ER, McGaghie WC, Wayne DB et al. Recommendations for Reporting Mastery Education Research in Medicine (ReMERM). Acad Med 2015; 90: 1509–1514

[13] Barsuk JH, Cohen ER, Feinglass J et al. Residents' procedural experience does not ensure competence: a research synthesis. J Grad Med Educ 2017; 9: 201–208

[14] Messick SA. Validity. 3 edn. New York: American Council on Education and Mac-Millan; 1989

[15] Downing SM YR, Haladyne TM, Axelson RD et al. Assessment in Health Professions Education. Routledge; 2009

[16] Cook DA, Lineberry M. Consequences validity evidence: evaluating the impact of educational assessments. Acad Med 2016; 91: 785–795

[17] Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Medical teacher 2012; 34: 960–992

[18] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016; 15: 155–163

[19] Jensen JT, Savran MM, Møller AM et al. Development and validation of a theoretical test in non-anaesthesiologist-administered propofol sedation for gastrointestinal endoscopy. Scand J Gastroenterol 2016; 51: 872–879

[20] Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. J Adv Nurs 2000; 32: 1008–1015

[21] Downing SM. Reliability: on the reproducibility of assessment data. Med Educ 2004; 38: 1006–1012

[22] Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ 2003; 37: 830–837

[23] Spada C, McNamara D, Despott EJ et al. Performance measures for small-bowel endoscopy: A European Society of Gastrointestinal Endoscopy (ESGE) Quality Improvement Initiative. United Europ Gastroenterol J 2019; 7: 614–641

[24] Fitts PM, Posner MI. Human performance. Grenwood Press; 1979

[25] Taylor JA, Ivry RB. The role of strategies in motor learning. Ann N Y Acad Sci 2012; 1251: 1–12

[26] Edwards WH. Motor learning and control: from theory to practice. Belmont: Wadsworth; 2011

[27] Park WG, Shaheen NJ, Cohen J et al. Quality indicators for EGD. Am J Gastroenterol 2015; 110: 60–71

[28] Konge L, Arendrup H, von Buchwald C et al. Using performance in multiple simulated scenarios to assess bronchoscopy skills. Respiration 2011; 81: 483–490

[29] Yudkowsky R, Park YS, Lineberry M et al. Setting mastery learning standards. Acad Medi 2015; 90: 1495–1500

[30] Kromann CB, Jensen ML, Ringsted C. The effect of testing on skills learning. Med Educ 2009; 43: 21–27

[31] Madsen ME, Konge L, Norgaard LN et al. Assessment of performance measures and learning curves for use of a virtual-reality ultrasound simulator in transvaginal ultrasound examination. Ultrasound Obstet Gynecol 2014; 44: 693–699

[32] Shirai Y, Yoshida T, Shiraishi R et al. Prospective randomized study on the use of a computer-based endoscopic simulator for training in esophagogastroduodenoscopy. J Gastroenterol Hepatol 2008; 23: 1046–1050

[33] Di Giulio E, Fregonese D, Casetti T et al. Training with a computer-based simulator achieves basic manual skills required for upper endoscopy: a randomized controlled trial. Gastrointest Endosc 2004; 60: 196–200

[34] Jirapinyo P, Abidi WM, Aihara H et al. Preclinical endoscopic training using a part-task simulator: learning curve assessment and determination of threshold score for advancement to clinical endoscopy. Surgical endoscopy 2017; 31: 4010–4015

[35] Preisler L, Svendsen MBS, Nerup N et al. Simulation-based training for colonoscopy: establishing criteria for competency. Medicine 2015; 94: e440

[36] Cold KM, Svendsen MBS, Bodtger U et al. Using structured progress to measure competence in flexible bronchoscopy. J Thorac Dis 2020; 12: 6797–6805

[37] Ritter EM, Taylor ZA, Wolf KR et al. Simulation-based mastery learning for endoscopy using the endoscopy training system: a strategy to improve endoscopic skills and prepare for the fundamentals of endoscopic surgery (FES) manual skills exam. Surg Endosc 2018; 32: 413–420