ORIGINAL RESEARCH

# Exploring Endoscopic Competence in Gastroenterology Training: A Simulation-Based Comparative Analysis of GAGES, DOPS, and ACE Assessment Tools

Faisal Wasim Ismail[1], Azam Afzal[1], Rafia Durrani[1], Rayyan Qureshi[1], Safia Awan[1], Michelle R Brown[2]

[1]Aga Khan University Karachi, Sind, Pakistan; [2]School of Health Professions, University of Alabama at Birmingham, Birmingham, AL, USA

Correspondence: Faisal Wasim Ismail, Department of Medicine, Aga Khan University, Stadium Road Karachi, Sind, 74800, Pakistan, Tel +922134863700, Fax +922134354678, Email faisal.ismail@aku.edu

**Purpose:** Accurate and convenient evaluation tools are essential to document endoscopic competence in Gastroenterology training programs. The Direct Observation of Procedural Skills (DOPS), Global Assessment of Gastrointestinal Endoscopic Skills (GAGES), and Assessment of Endoscopic Competency (ACE) are widely used validated competency assessment tools for gastrointestinal endoscopy. However, studies comparing these 3 tools are lacking, leading to lack of standardization in this assessment. Through simulation, this study seeks to determine the most reliable, comprehensive, and user-friendly tool for standardizing endoscopy competency assessment.

**Methods:** A mixed-methods quantitative-qualitative approach was utilized with sequential deductive design. All nine trainees in a gastroenterology training program were assessed on endoscopic procedural competence using the Simbionix Gi-bronch-mentor high-fidelity simulator, with 2 faculty raters independently completing the 3 assessment forms of DOPS, GAGES, and ACE. Psychometric analysis was used to evaluate the tools' reliability. Additionally, faculty trainers participated in a focused group discussion (FGD) to investigate their experience in using the tools.

**Results:** For upper GI endoscopy, Cronbach's alpha values for internal consistency were 0.53, 0.8, and 0.87 for ACE, DOPS, and GAGES, respectively. Inter-rater reliability (IRR) scores were 0.79 (0.43–0.92) for ACE, 0.75 (−0.13–0.82) for DOPS, and 0.59 (−0.90–0.84) for GAGES. For colonoscopy, Cronbach's alpha values for internal consistency were 0.53, 0.82, and 0.85 for ACE, DOPS, and GAGES, respectively. IRR scores were 0.72 (0.39–0.96) for ACE, 0.78 (−0.12–0.86) for DOPS, and 0.53 (−0.91–0.78) for GAGES. The FGD yielded three key themes: the ideal tool should be scientifically sound, comprehensive, and user-friendly.

**Conclusion:** The DOPS tool performed favourably in both the qualitative assessment and psychometric evaluation to be considered the most balanced amongst the three assessment tools. We propose that the DOPS tool be used for endoscopic skill assessment in gastroenterology training programs. However, gastroenterology training programs need to match their learning outcomes with the available assessment tools to determine the most appropriate one in their context.

**Keywords:** simulation, training, competence, endoscopy, gastroenterology

## Introduction

Gastrointestinal endoscopy has been abundantly used as a minimally invasive diagnostic and therapeutic procedure for a plethora of illnesses since its advent by Adolf Kussmaul in 1868.[1] Due to widespread use, growing demand, and increasing requirements of trainees, the role of assessment tools and simulation has become important in assessing competency of endoscopic skills.[2–5]Previously, a set number of procedures were used as a benchmark to establish accreditation and granting privileges to conduct such procedures.[6,7] This number also varied significantly between different organizations and countries. Recently, the American Society for Gastrointestinal Endoscopy (ASGE) and the Joint Advisory Group for Gastrointestinal Endoscopy (JAG) have both made recommendations for establishment and

**75**

maintenance of quality and standards of endoscopic training and intervention in the United States (US) and the United Kingdom (UK), respectively.[8–10]

Various clinical assessment tools have been developed and validated to ensure competency in gastroenterology trainees. Two such prominent and validated tools include Direct Observation of Procedural Skills (DOPS) used by JAG for Gastroenterology in the UK, and the Global Assessment of Gastrointestinal Endoscopic Skills (GAGES) used by the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) in the US.[11,12] However, these tools focus primarily on motor skills associated with endoscopy with little emphasis on cognitive skills.

The ASGE developed another tool namely Assessment of Competency in Endoscopy (ACE) that assesses both motor and cognitive skills in a comprehensive manner.[13] ACE has been conceived using the Mayo Colonoscopy Skills Assessment Tool (MCSAT) as a platform whilst bridging its limitations. Because it is a common tool that is used extensively to assess competence in GI in parts of the world,[14–16] we considered it an important one to be included in this study.

There is a dearth of literature comparing the performance of the aforementioned assessment tools, despite a wealth of well-documented research on their individual validation. Consequently, there is a lack of standardization of endoscopy competence assessment across training programs, leading to heterogeneity between endoscopists and training programs as to which assessment tool is superior.[14,17,18]

Thus, it is imperative to implement a standardized and structured process for evaluating endoscopic skill competence to facilitate accurate trainee comparisons.[13,19,20] With multiple assessment tools available, it is essential to select the most reliable, discriminative, and user-friendly tool, for it to be used as a standard tool of assessment.

Simulation as an educational instrument allows trainees to perform in a controlled environment with different levels of difficulty while maintaining patient safety. Various studies have shown it to boost trainee confidence, improve patient outcomes and reduce procedure related discomfort.[3,21,22] It has been a widely and extensively used modality in endoscopy training.[23,24] The standard cases and the identical nature of difficulty lend simulation very well to assess competence in a cohort of trainees with similar baseline skills.[25]

This study's framework was based on the Accreditation Council for Graduate Medical Education (ACGME) Advisory Committee on Educational Outcome Assessment's framework for assessing the quality of an assessment measure. The ACGME defines six criteria for evaluating standards, including reliability, validity, ease of use, required resources, ease of interpretation, and educational impact.[26] Since each of the three tools are independently validated and reliable, we chose to employ ease of use, required resources, and ease of interpretation in our study.

The objective of our study is to compare three different assessment tools to identify the tool that most accurately reflects competence in endoscopic skills.

## Methodology

A mixed methods quantitative-qualitative approach was employed, specifically utilizing a deductive-sequential design where the primary component was quantitative, and the secondary component was qualitative.

## Materials

Endoscopic Assessment Tools:

1. Direct observation of procedural skills (DOPS): Attached as Supplementary Materials 1 and 2 for EGD and colonoscopy, respectively. It uses a numerical scale as follows: Score 1 (requiring maximal supervision), Score 2 (requiring significant supervision), Score 3 (requiring minimal supervision), Score 4 (competent for unsupervised practice).
2. Global Assessment of Gastrointestinal Endoscopic Skills (GAGES): Attached as Supplementary Material 3 and 4 for EGD and colonoscopy, respectively. It uses a similar scale ranging from 1 (unsatisfactory) to 5 (competent).
3. Assessment of Competency in Endoscopy (ACE) tool: It uses a 4-point grading scale defining the continuum from novice (1) to competent (4).[13]

## Quantitative Methodology

Participants were trainees in the Gastroenterology training program at the Aga Khan University Hospital (AKUH), Karachi Pakistan. The AKUH is a 750-bed private tertiary care teaching hospital serving the needs of the city of Karachi, and the country at large. The Gastroenterology training program at the AKUH is a 3-year program. As part of their training, the trainees have defined competency goals at each level of training, for which they are required to achieve competence in particular skills by the end of that training year.

The gastroenterology training curriculum utilizes simulation to develop competence in trainees. As per this curriculum, they practice endoscopic skills on the Simbionix GI-BRONCH-mentor high fidelity simulator[27] until they achieve competence as per the simulator generated report for that procedure. They can then start supervised procedures on patients with faculty.

The sample included all 9 gastroenterology trainees in the GI training program during the year 2020–21, which constituted 100% of the population under study.

For the study, the trainee had to be deemed competent as per the simulator report in upper GI endoscopy (EGD) and colonoscopy. Once the required cohort was ready, they were taken through pre-simulation briefing as per the Standards of Best Practice for simulation briefing,[28] introducing all aspects of the simulation and explaining the learning objectives in detail as well as the tasks expected from them to perform. PGY1 trainees performed 2 upper GI endoscopies, while trainees in PGY2 and PGY3 performed 2 upper GI endoscopies and 2 colonoscopies. PGY1 trainees were not made to perform colonoscopies as they had not learnt how to perform them at the time of the study.

The cases were pre-decided by faculty, were not known to the trainee, and were diagnostic in nature. The cases for all the trainees were different, but of the same difficulty level as per the in-built modules of Simbionix GI-Bronch-Mentor high-fidelity simulator. These modules in software version SI-2825 comprise cases with similar findings and difficulty for learners to practice and were utilized for the study. For EGD, case 1 from module 1 was used, which was about a 30-year -old male with epigastric pain referred for endoscopy. For colonoscopy, case 1 from module 1 was used, which was about a 30-year-old female with irregular bowels referred for colonoscopy.

A simulated patient was used to evaluate the consent process, and a nurse assisted the participant in the conduct of the simulated procedure by informing the participant about the vital signs of the patient and assisting in taking a biopsy sample. Two faculty raters directly observed the trainee's performing the procedures. Once a trainee performed a procedure, 2 faculty independently, and without consultation, completed all three of the assessment tools. This structure was followed for each procedure. Prior to starting, structured rater training was done. Rater training was held in a half-day workshop for all the raters. A didactic lecture on rater error training was conducted to familiarize trainers about common rating errors, followed by a Frame-Of-Reference (FOR) training[29] using videos where the group discussed discrepancies between individual raters with a facilitator.

## Qualitative Methodology

Though psychometrics is important for providing validity and reliability evidence for assessment tools, the factors which often effective implementation of the assessment rely on its feasibility (cost, time and resource effectiveness) and acceptability (by faculty and students). It is for this reason that

> educators need information regarding not only the trustworthiness of scores, but also the logistics and practical issues such as cost, acceptability, and feasibility that arise during test implementation and administration.[30]

Therefore, a qualitative component was added to the study to provide evidence for ease of use, resources required, and ease of interpretation. For this, a focus group discussion (FGD) was considered suitable[31,32] and, hence, conducted for one hour to collect data about the utilization and implementation of the different assessment tools. Gastroenterology faculty (n=6) who assessed trainee performances, took part in this FGD. The purpose of the focused group was to gather information from the faculty about assessing trainee performance using the different tools. The prompt questions (attached as Supplementary Material 5) used for the FGD were designed to probe raters about their experiences using the multiple assessment tools. These questions explored rater experience with the tool and ease of interpretation which could help the authors qualitatively understand if the rater training that was conducted was helpful in minimizing biases.

It also helps authors understand raters' opinion about feasibility and acceptability, which are important qualitative characteristics of simulation assessment tools.[33] The FGD was recorded after all participants provided informed consent. The audio recordings were transcribed to a written document by an author who was not part of the simulation sessions or initial study design but was only provided the audio recording. The first author validated the transcribed document, while all authors generated the themes after reviewing the transcription.

## Data Collection & Analysis

Participant scores were collected individually. Component and total scores of the DOPS, GAGES, and ACE skills were calculated individually for each procedure. An analysis was conducted where the mean and standard deviation were computed for each scale. The total score of each tool was stratified by level of training (ie, PGY1, PGY2, and PGY3) using one-way analysis of variance.

To assess internal consistency, Cronbach's alpha was computed for each scale. Additionally, inter-rater reliability (IRR) was evaluated using the Intraclass Correlation. Alpha and IRR calculations were done with all trainees included in the analysis. All these analyses were performed using SPSS version 25.

Psychometric Analysis: Reliability analysis (internal consistency and inter-rater reliability) was calculated for each tool.

## Reliability Analysis

(a) Internal consistency describes the extent to which all the items in a test measure the same concept or construct and, hence, it is connected to the inter-relatedness of the items within the test.[34] Cronbach's Alpha is the measure of internal consistency for ordinal and interval data.[35,36] The choice of cut-off values varies based on the type of examination: a threshold of 0.90 is recommended for high-stakes exams, while the range of 0.80–0.89 is considered acceptable for moderate tests, and 0.70–0.79 is deemed acceptable for lower-stakes tests.

(b) Inter-rater reliability (IRR) refers to the level of agreement achieved by independent assessors who rate the same performance. Intraclass correlation is used for ordinal and interval data[37] and is interpreted as the proportion of variance in the ratings of the phenomenon being rated.[35] Therefore, intraclass correlation of the scores given by different raters to the same trainee can be analyzed to provide evidence for inter-rater reliability.[35]

## Qualitative Analysis

The FGD was recorded and transcribed for data collection. After member checking, content analysis was done to determine each tool's ease of use, ease of interpretation and cognitive load.

This study was in compliance with the World Medical Association Declaration of Helsinki[38] and was approved by the Aga Khan University ethical review committee. We strictly monitored and maintained data confidentiality. Each trainee was given a unique identity and trainee identifiers were kept anonymous.; only the research team had access to this material. The data was stored as per good clinical practice guidelines. Written informed consent was taken from all participants prior to study commencement, and the informed consent included publication of anonymized responses.

## **Results**

### Demographic Data

The faculty raters were 6 in number. The trainees were 9 in number (2 PGY1, 3 PGY2 and 4 PGY3 trainees), and all of them met the inclusion criteria.

Participant scores for each tool

Table 1 shows a summary of average scores for each tool by level of training.

Quantitative analysis for upper GI endoscopy – esophagogastroduodenoscopy (EGD)

Table 2 shows the quantitative analysis of each tool for upper GI endoscopy, with Cronbach's Alpha for internal consistency of the tools and Inter-Rater Reliability.

Quantitative analysis for colonoscopy

**Table 1** Summary of Scores for Each Tool According to the Level of Training

| Tool | PGY1 | PGY2 | PGY3 | p-value |
|------|------|------|------|---------|
| ACE | 19.8 ± 1.3 | 24 ± 1.2 | 26.3 ± 2.6 | <0.001 |
| DOPS | 59.2 ± 1.4 | 60 ± 2.4 | 56.8 ± 5.1 | 0.06 |
| GAGES | 18.7 ± 0.5 | 24.6 ± 0.5 | 24.1 ± 0.9 | <0.001 |

**Table 2** Quantitative Analysis for Upper GI Endoscopy – Esophagogastroduodenoscopy (EGD)

| Tool | Cronbach's Alpha | IRR (CI) |
|------|------------------|----------|
| ACE | 0.53 | 0.79 (0.43–0.92) |
| DOPS | 0.8 | 0.75 (−0.13–0.82) |
| GAGES | 0.87 | 0.59 (−0.90–0.84) |

**Table 3** Quantitative Analysis for Colonoscopy

| Tool | Cronbach's Alpha | IRR (CI) |
|------|------------------|----------|
| ACE | 0.53 | 0.72 (0.39–0.96) |
| DOPS | 0.82 | 0.78 (−0.12–0.86) |
| GAGES | 0.85 | 0.53 (−0.91–0.78) |

Table 3 shows the quantitative analysis of each tool for colonoscopy, with Cronbach's Alpha for internal consistency of the tools and Inter-Rater Reliability.

## Qualitative Analysis

A series of open-ended questions were asked regarding the ease of use of the tools, the strengths and areas of improvement of each tool, the comprehensiveness of each tool, and the resources that were required to complete the tool. Analysis of the transcript resulted in three major themes emerging regarding the tools, namely that the best tool needs to be: scientifically sound, comprehensive and user-friendly. The next section will describe these themes in further detail and provide quotations that most accurately depict the themes and dimensions emerging from the FGD.

## Scientifically Sound Tool

We asked the trainers which tool in their opinion was the most complete, in terms of the robustness of assessment and utility of use. All universally said that the most important was that the tool be valid and reliable for the assessment, and that no compromise could be made on its scientific credentials.

Tools need to be valid and reliable (All trainers)

No point in having a tool that is not scientifically reliable-the assessment will be completely pointless (Trainer 2)

On asking about further qualities, variation was noted between the trainers, where some felt that the increased detail in the ACE is essential, while others felt that the same detail made it overwhelming.

ACE- Very detailed, breaks down the skill into many parts (trainer 4). Assesses knowledge-that's good, but we are assessing endoscopic skill here (trainer 2) Very detailed and complete-but perhaps difficult to do on a formative basis regularly (trainer 5)

Trainers also felt that the best tool needs to assess not only the endoscopic procedure itself but sedation, safety, consent and communication with the patient and para-medical staff before and during the procedure and post-procedure tasks such as report generation, and capturing of images as well.

Tools need to assess not only the psychomotor skills, but the procedure in totality-the pre and post procedure experience as well (trainer 1)

DOPS- Covers most aspects of a good assessment (trainer 2) It's a good complete tool-does what it's supposed to do (trainer 4)

GAGES- Isolates the procedure from the other aspects (trainer 3) Gives you a broad assessment (trainer 5) Too procedure focused only (trainer 6)

## User-Friendly Tool

Trainers said that the most important thing for them to assess regarding the utility of the tool was its ease of use. This was one of the most important factors that word dictate whether they would continue to use it in the future. They defined ease of use as the form not taking long to fill, not very wordy, with an open layout and easy readability with regard to font size and type.

Tools need to be easy to use, quick to fill, assess all major skills (trainer 2)

The GAGES was super-quick. DOPS was next-the ACE was good, but just too cumbersome (trainer 5)

You have to see which tools can be used repeatedly, so that they don't become overwhelming and time-consuming (trainer 1)

You have to look for a happy medium-not too long, but not too short, and easy to understand (Trainer 3 and 4)

## Comprehensive Tool

The trainers were asked their opinion regarding the resources that were required to complete the tools. All of them felt that there was no difference in material resources. However, the time that was required to fill the tools was more for the ACE as compared to the GAGES and DOPS.

I didn't require anything different in completing one tool or another (trainer 4)

Is time a resource? -it is, right? Then the ACE is definitely more resource intensive than the other 2 (trainer 6)

## Pros and Cons of Each Tool

The trainers were asked to discuss the advantages and disadvantages of every tool in detail. This was asked in turn from all six trainers. General comments about any tool were not adverse. All the trainers thought that all three tools were generally very good, assessed the skills in detail, and were good to use on their own.
  However, they all had their differences which made the overall experience of using them different for the trainers.

GAGES- Quick and to the point, but too brief (trainer 1)

DOPS- Reasonably quick, but takes a bit of initial understanding (trainer 3) Nice, open layout (trainer 5) Measures what it's supposed to (trainer 6)

ACE- Detailed assessment, but time consuming (trainer 4) lengthy (trainer 6) Overwhelming (trainer 3) way too much detail-difficult to do regularly as formative assessment (trainer 5)

# Discussion

The traditional approach to teaching endoscopic procedures involves the apprenticeship model, which relies on procedure numbers and training time. However, this approach is risky and unpredictable, potentially endangering patient safety. Furthermore, this training method lacks standardization and is subject to variation based on the trainer, trainee, and patient. To ensure a consistent and objective evaluation of competence, training must be standardized and structured, with frequent formative assessments.[39] In this study, we aimed to identify, through a scientifically rigorous and structured approach, the most effective assessment tool for this purpose.

Our results were in line with the values obtained in the initial validity studies of these tools. For ACE, the reliability coefficient was calculated for its individual components and reported as 0.79 for cognitive part and 0.88 for the motor part.[40] GAGES had Cronbach's alpha value of 0.89 for upper GI endoscopy and 0.95 for colonoscopy.[12] Its IRR was 0.96 for upper GI endoscopy and 0.97 for colonoscopy.[12] DOPS had a Cronbach's alpha value of 0.95[41] and reliability of 0.81.[11]

A key concept in the interpretation of internal consistency is that a low alpha value might result from a limited number of questions, weak correlations among items, or diverse underlying constructs.[34] Conversely, an excessively elevated alpha could indicate item redundancy, where certain items essentially examine the same concept but are phrased differently.[34] Our study determined that in EGD skill assessment, the highest internal consistency was for the GAGES, followed by the DOPS tool. As per the key interpretation concept mentioned above, the GAGES tool would be expected to score higher, given the fact that it is a 5-point scale and therefore a wider range, resulting in a greater standard deviation. In contrast, DOPS and ACE each are 4-point scales, hence having ranges narrower than GAGES, resulting in lower standard deviation and scores. Additionally, ACE has diverse underlying constructs, resulting in lower alpha values. The tools performed in a similar fashion for colonoscopy assessment as well. Trends remained the same as EGD for both analyses in colonoscopy in all three tools.

As discussed earlier, all 3 tools are very robust and can be used safely in the assessment of GI endoscopic skills. The GAGES tool is easy to administer, is consistent and meets high standards of reliability and validity.[12] The DOPS tool is a very good in-training competency assessment tool,[42] while the ACE tool is also very sound and can meaningfully monitor fellows' progress throughout the entire training continuum.[43]

GAGES is designed to measure technical skill,[12] which it does very well. However, in our study, raters felt that a good endoscopic skill assessment tool should include the pre- and post-technical skill attributes as well, such as communication with the patient and staff, consent, and recognition of findings and complications. Given this, and the relative subjectivity in the scoring, the GAGES tool is our study was thought to be a good tool only if procedural assessment was required but not if the goal was training an endoscopy learner in the overall aspects that make up the procedure.

The ACE tool was extremely detailed and reduced subjectivity the most, thus having the highest IRR. However, in the FGD, raters felt that the tool was very extensive and detailed, and while all felt that may be a good thing, they felt it did not lend itself to repeated use. A few raters felt that there were so many entries to make, which took away from the time they would like to give to the trainee in providing feedback. Some felt that the ACE may serve well for advanced trainees as summative assessment, but for formative assessment they would prefer a briefer tool.

The DOPS tool scored well in the reliability analysis. It had favorable reviews in the FGD as well. Raters felt it did not ignore the pre- and post-procedure steps of an endoscopic procedure, was not time-consuming, and could be easily performed as formative and summative assessment. The scale was easy to understand and seemed to have the best mix of the affective, knowledge and technical attributes required for effective skills assessment of a gastroenterology trainee. These findings are also supported by Siau et al in their study about nationwide evidence on validity and reliability evidence to assess DOPS as an assessment tool[42] and in the study on pediatric EGD.[44]

Our study had main strengths, discussed as follows. To minimize bias and reduce sources of error in our study, we ensured that each trainee's performance evaluation was rated by at least two individuals. The assessment tools we used for scoring the trainees' performance were all validated tools identified from literature.[45]

We utilized a standardized patient to evaluate the consent taking process, and a nurse for procedural assistance. We did this so that all facets of the tools could be evaluated, to avoid stereotyping and maintain fidelity, in keeping with best practices in simulation.

Prior to conducting the performance assessment, raters were trained through a discussion of the tools. This training, combined with independent assessments by the two raters, helped to establish the inter-rater reliability demonstrated in our study results.

The primary limitation observed in the study is the small sample size, despite the inclusion of all trainees. This was compensated for by the qualitative aspect of the study which had data from the FGDs with all the raters which, in turn, may have brought a subjective bias into the study. Additionally, the study was conducted exclusively within a single institution.

The results of our study have demonstrated the DOPS tool to be the most consistent assessment tool out of the three, in terms of internal consistency and inter-rater reliability as well as the themes identified from FGDs. However, our results also present an array of options for choosing the best tool based on the nature of the assessment being done. GAGES would be best applicable to situations in which only the procedural skill is being assessed. DOPS would be the best for overall procedural competence in a comprehensive yet compact manner, including both the knowledge and skill of the trainee, whereas ACE is most likely to have utility in summative examinations in which detailed assessment is being done for all aspects of procedural competence in depth.

These results have important implications for teaching and training in GI training programs. It will benefit trainees by identifying the best tool to provide them with vital feedback for learning. Selection of an appropriate assessment tool will help faculty by making scoring easier, more objective and efficient. This will improve training programs by ensuring standardization of assessment through valid and reliable tools to ensure high-quality training. Additional studies could be conducted to investigate the effect of training using different assessment tools, by examining their correlation with the clinical performance and quality outcome indicators of trainees in real-world clinical settings.

## Conclusion

Our study demonstrated that all three assessment tools were found to have their advantages and drawbacks, and it would be inappropriate to pick one as the best for all contexts. In our study, the DOPS tool performed favorably in the qualitative assessment, and performed adequately in the psychometric evaluation to warrant being considered the most balanced of the tools. Therefore, we propose that the DOPS tool be used for endoscopic skill assessment in gastro-enterology training programs whenever the setting resembles conditions portrayed in our study. However, we propose that gastroenterology training programs match their learning outcomes with the available assessment tools to decide which may be the most appropriate one for their need and context.

## Disclosure

All authors report no conflicts of interest in this work.

## References

1. Sivak MV. Gastrointestinal endoscopy: past and future. *Gut*. 2005;55(8):1061–1064. doi:10.1136/gut.2005.086371
2. Bittner JG, Marks JM, Dunkin BJ, Richards WO, Onders RP, Mellinger JD. Resident training in flexible gastrointestinal endoscopy: a review of current issues and options. *J Surg Educ*. 2007;64(6):399–409. doi:10.1016/j.jsurg.2007.07.006
3. Haycock A, Koch AD, Familiari P, et al. Training and transfer of colonoscopy skills: a multinational, randomized, blinded, controlled trial of simulator versus bedside training. *Gastrointest Endosc*. 2010;71(2):298–307. doi:10.1016/j.gie.2009.07.017
4. Shenbagaraj L, Thomas-Gibson S, Stebbing J, et al. Endoscopy in 2017: a national survey of practice in the UK. *Frontline Gastroenterol*. 2019;10 (1):7–15. doi:10.1136/flgastro-2018-100970
5. Matharoo M, Thomas-Gibson S. Safe endoscopy. *Frontline Gastroenterol*. 2017;8(2):86–89. doi:10.1136/flgastro-2016-100766
6. Eisen GM, Baron TH, Dominitz JA, et al. Methods of granting hospital privileges to perform gastrointestinal endoscopy. *Gastrointest Endosc*. 2002;55(7):780–783. doi:10.1016/S0016-5107(02)70403-3
7. Society of American Gastrointestinal Endoscopic Surgeons (SAGES). Granting of privileges for gastrointestinal endoscopy by surgeons. *Surg Endosc*. 1998;12(4):381–382. doi:10.1007/s004649900685
8. Faulx AL, Lightdale JR, Acosta RD, et al. Guidelines for privileging, credentialing, and proctoring to perform GI endoscopy. *Gastrointest Endosc*. 2017;85(2):273–281. doi:10.1016/j.gie.2016.10.036

9. Beg S, Banks M. Improving quality in upper gastrointestinal endoscopy. *Frontline Gastroenterol*. 2022;13(3):184–185. doi:10.1136/flgastro-2021-101946

10. Siau K, Pelitari S, Green S, et al. JAG consensus statements for training and certification in colonoscopy. *Frontline Gastroenterol*. 2023;14(3):201–221. doi:10.1136/flgastro-2022-102260

11. Barton JR, Corbett S, Van Der Vleuten CP. The validity and reliability of a direct observation of procedural skills assessment tool: assessing colonoscopic skills of senior endoscopists. *Gastrointest Endosc*. 2012;75(3):591–597. doi:10.1016/j.gie.2011.09.053

12. Vassiliou MC, Kaneva PA, Poulose BK, et al. Global assessment of gastrointestinal endoscopic skills (GAGES): a valid measurement tool for technical skills in flexible endoscopy. *Surg Endosc*. 2010;24(8):1834–1841. doi:10.1007/s00464-010-0882-8

13. Fried GM, Marks JM, Mellinger JD, Trus TL, Vassiliou MC, Dunkin BJ. ASGE's assessment of competency in endoscopy evaluation tools for colonoscopy and EGD. *Gastrointest Endosc*. 2014;80(2):366–367. doi:10.1016/j.gie.2014.03.019

14. Sedlack RE, Coyle WJ, Research Group ACE. Assessment of competency in endoscopy: establishing and validating generalizable competency benchmarks for colonoscopy. *Gastrointest Endosc*. 2016;83(3):516–523.e1. doi:10.1016/j.gie.2015.04.041

15. Sabrie N, Khan R, Seleq S, et al. Global trends in training and credentialing guidelines for gastrointestinal endoscopy: a systematic review. *Endosc Int Open*. 2023;11(02):E193–E201. doi:10.1055/a-1981-3047

16. Miller AT, Sedlack RE, Sedlack RE, et al. Competency in esophagogastroduodenoscopy: a validated tool for assessment and generalizable benchmarks for gastroenterology fellows. *Gastrointest Endosc*. 2019;90(4):613–620.e1. doi:10.1016/j.gie.2019.05.024

17. Scarallo L, Russo G, Renzo S, Lionetti P, Oliva S. A journey towards pediatric gastrointestinal endoscopy and its training: a narrative review. *Front Pediatr*. 2023;11:1201593. doi:10.3389/fped.2023.1201593

18. Ekkelenkamp VE, Koch AD, De Man RA, Kuipers EJ. Training and competence assessment in GI endoscopy: a systematic review. *Gut*. 2016;65(4):607–615. doi:10.1136/gutjnl-2014-307173

19. Bjorkman DJ, Popp JW. Measuring the quality of endoscopy. *Gastrointest Endosc*. 2006;63(4):S1–S2. doi:10.1016/j.gie.2006.02.022

20. Siau K, Hawkes ND, Dunckley P. Training in Endoscopy. *Curr Treat Options Gastroenterol*. 2018;16(3):345–361. doi:10.1007/s11938-018-0191-1

21. Higham H, Baxendale B. To err is human: use of simulation to enhance training and patient safety in anaesthesia. *Br J Anaesth*. 2017;119:i106–i114. doi:10.1093/bja/aex302

22. Wheelan SA, Burchill CN, Tilin F. The link between teamwork and patients' outcomes in intensive care units. *Am J Crit Care Off Publ Am Assoc CritCare Nurses*. 2003;12(6):527–534.

23. Khan R, Plahouras J, Johnston BC, Scaffidi MA, Grover SC, Walsh CM. Virtual reality simulation training for health professions trainees in gastrointestinal endoscopy. Cochrane colorectal cancer group. *Cochrane Database Syst Rev*. 2018;2018(8). doi:10.1002/14651858.CD008237.pub3

24. Finocchiaro M, Cortegoso Valdivia P, Hernansanz A, et al. Training simulators for gastrointestinal endoscopy: current and future perspectives. *Cancers*. 2021;13(6):1427. doi:10.3390/cancers13061427

25. Neily J, Mills PD, Young-Xu Y, et al. Association between implementation of a medical team training program and surgical mortality. *JAMA*. 2010;304(15):1693. doi:10.1001/jama.2010.1506

26. Swing SR, Clyman SG, Holmboe ES, Williams RG. Advancing resident assessment in graduate medical education. *J Grad Med Educ*. 2009;1(2):278–286. doi:10.4300/JGME-D-09-00010.1

27. Simbionix. GI Mentor | simbionix. Simbionix | to advance clinical performance; 2020. Available from: https://simbionix.com/simulators/gi-mentor/. Accessed January 16, 2024.

28. McDermott DS, Ludlow J, Horsley E, Meakim C. Healthcare simulation standards of best practice TM prebriefing: preparation and briefing. *Clin Simul Nurs*. 2021;58:9–13. doi:10.1016/j.ecns.2021.08.008

29. Feldman M, Lazzara EH, Vanderbilt AA, DiazGranados D. Rater training to support high-stakes simulation-based assessments. *J Contin Educ Health Prof*. 2012;32(4):279–286. doi:10.1002/chp.21156

30. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. *Adv Simul Lond Engl*. 2016;1:31. doi:10.1186/s41077-016-0033-y

31. Schuster RC, Brewis A, Wutich A, et al. Individual interviews versus focus groups for evaluations of international development programs: systematic testing of method performance to elicit sensitive information in a justice study in Haiti. *Eval Program Plann*. 2023;97:102208. doi:10.1016/j.evalprogplan.2022.102208

32. Baillie L. Exchanging focus groups for individual interviews when collecting qualitative data. *Nurse Res*. 2019;27:15–20. doi:10.7748/nr.2019.e1633

33. Nunnink L, Foot C, Venkatesh B, et al. High-stakes assessment of the non-technical skills of critical care trainees using simulation: feasibility, acceptability and reliability. *Crit Care Resusc J Australas Acad Crit Care Med*. 2014;16(1):6–12.

34. Tavakol M, Dennick R. Making sense of Cronbach's alpha. *Int J Med Educ*. 2011;2:53–55. doi:10.5116/ijme.4dfb.8dfd

35. Holcomb ZC, Cox KS. *Interpreting Basic Statistics: A Workbook Based on Excerpts from Journal Articles*. 18th. Routledge; 2017. doi:10.4324/9781315225647

36. Gadermann A, Guhn M, Zumbo BD. Ordinal Alpha. In: Michalos AC editor. *Encyclopedia of Quality of Life and Well-Being Research*. Springer Netherlands; 2014:4513–4515. doi:10.1007/978-94-007-0753-5_2025

37. Hallgren KA. Computing Inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol*. 2012;8(1):23–34. doi:10.20982/tqmp.08.1.p023

38. World Medical Association. World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*. 2013;310(20):2191–2194. doi:10.1001/jama.2013.281053

39. Kim Y, Lee JH, Lee GH, et al. Simulator-based training method in gastrointestinal endoscopy training and currently available simulators. *Clin Endosc*. 2023;56(1):1–13. doi:10.5946/ce.2022.191

40. Sedlack RE. The mayo colonoscopy skills assessment tool: validation of a unique instrument to assess colonoscopy skills in trainees. *Gastrointest Endosc*. 2010;72(6):1125–1133, 1133.e1–3. doi:10.1016/j.gie.2010.09.001

41. Siau K, Hodson J, Anderson JT, et al. PWE-111 Impact of the jag basic skills in colonoscopy course on trainee performance. *Posters*. BMJ Publishing Group Ltd and British Society of Gastroenterology; 2019:A255–A256. doi:10.1136/gutjnl-2019-BSGAbstracts.482

42. Siau K, Crossley J, Dunckley P, et al. Direct observation of procedural skills (DOPS) assessment in diagnostic gastroscopy: nationwide evidence of validity and competency development during training. *Surg Endosc*. 2020;34(1):105–114. doi:10.1007/s00464-019-06737-7

43. Walsh CM, Ling SC, Khanna N, et al. Gastrointestinal endoscopy competency assessment tool: reliability and validity evidence. *Gastrointest Endosc.* 2015;81(6):1417–1424.e2. doi:10.1016/j.gie.2014.11.030

44. Siau K, Levi R, Howarth L, et al. Validity evidence for direct observation of procedural skills in paediatric gastroscopy. *J Pediatr Gastroenterol Nutr.* 2018;67(6):e111–e116. doi:10.1097/MPG.0000000000002089

45. Khan R, Zheng E, Wani SB, et al. Colonoscopy competence assessment tools: a systematic review of validity evidence. *Endoscopy.* 2021;53 (12):1235–1245. doi:10.1055/a-1352-7293