Assessment of competency in endoscopy: establishing and validating generalizable competency benchmarks for colonoscopy

Robert E. Sedlack, MD, MHPE,[1] **Walter J. Coyle, MD,**[2] **ACE Research Group**

Rochester, Minnesota; La Jolla, California, USA

Background and Aims: The Mayo Colonoscopy Skills Assessment Tool (MCSAT) has previously been used to describe learning curves and competency benchmarks for colonoscopy; however, these data were limited to a single training center. The newer Assessment of Competency in Endoscopy (ACE) tool is a refinement of the MCSAT tool put forth by the Training Committee of the American Society for Gastrointestinal Endoscopy, intended to include additional important quality metrics. The goal of this study is to validate the changes made by updating this tool and establish more generalizable and reliable learning curves and competency benchmarks for colonoscopy by examining a larger national cohort of trainees.

Methods: In a prospective, multicenter trial, gastroenterology fellows at all stages of training had their core cognitive and motor skills in colonoscopy assessed by staff. Evaluations occurred at set intervals of every 50 procedures throughout the 2013 to 2014 academic year. Skills were graded by using the ACE tool, which uses a 4-point grading scale defining the continuum from novice to competent. Average learning curves for each skill were established at each interval in training and competency benchmarks for each skill were established using the contrasting groups method.

Results: Ninety-three gastroenterology fellows at 10 U.S. academic institutions had 1061 colonoscopies assessed by using the ACE tool. Average scores of 3.5 were found to be inclusive of all minimal competency thresholds identified for each core skill. Cecal intubation times of less than 15 minutes and independent cecal intubation rates of 90% were also identified as additional competency thresholds during analysis. The average fellow achieved all cognitive and motor skill endpoints by 250 procedures, with >90% surpassing these thresholds by 300 procedures.

Conclusions: Nationally generalizable learning curves for colonoscopy skills in gastroenterology fellows are described. Average ACE scores of 3.5, cecal intubation rates of 90%, and intubation times less than 15 minutes are recommended as minimal competency criteria. On average, it takes 250 procedures to achieve competence in colonoscopy. The thresholds found in this multicenter cohort by using the ACE tool are nearly identical to the previously established MCSAT benchmarks and are consistent with recent gastroenterology training recommendations but far higher than current training requirements in other specialties. (Gastrointest Endosc 2016;83:516-23.)

In 2011, data defining the learning curves and competency benchmarks of gastroenterology fellows learning the procedure of colonoscopy were reported by using the Mayo Colonoscopy Skills Assessment Tool (MCSAT).[1] The use of this tool and the established benchmarks allow for continuous assessment of the trainee's progression toward competence along the entire training continuum from novice to competence.

Abbreviations: ACE, Assessment of Competency in Endoscopy; ASGE, American Society for Gastrointestinal Endoscopy; CI, confidence interval; MCSAT, Mayo Colonoscopy Skills Assessment Tool; NAS, Next Accreditation System.

DISCLOSURE: *This project was funded by an unrestricted grant from the Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, Minnesota. The Mayo Clinic holds the rights to the Assessment of Competency in Endoscopy tool and the Mayo Colonoscopy Skills Assessment Tool (© Mayo Foundation for Medical Education and Research). Dr Sedlack has partial interest in the tools mentioned. Dr Coyle disclosed no financial relationships relevant to this publication.*

0016-5107/$36.00
http://dx.doi.org/10.1016/j.gie.2015.04.041

Received February 12, 2015. Accepted April 26, 2015.

Current affiliations: Mayo Clinic, Rochester, Minnesota (1), Scripps Clinic, La Jolla, California (2), USA.

Reprint requests: Robert Sedlack, MD, MHPE, Mayo Clinic, 200 First Street SW, Mayo 9-E, Rochester, MN 55905.

Despite the strengths of the MCSAT, refinements were needed, such as a metric assessing fine tip control and a means to calculate polyp detection rates. The Training Committee of the American Society for Gastrointestinal Endoscopy (ASGE) took up the task of making these additions to this tool with the results being the Assessment of Competency in Endoscopy (ACE) tool for colonoscopy reported in a recent white paper published in 2014.[2] Additionally, the MCSAT learning curves previously reported are based on data from a single training program. Whether these same trends are applicable to other fellowship centers has not been studied. The goals of this study are to validate the newly added metrics of the ACE tool and revalidate the tool as a whole. Additionally, this study should define generalizable national learning curves and competency thresholds for colonoscopy by using the ACE scoring system.

METHODS

Study design

This was a prospective, multicenter study validating the ACE tool and assessment of trainee learning curves and competency endpoints in colonoscopy. This study was performed across 10 gastroenterology training centers across the United States, during the academic year July 2013 through June 2014. Cases were performed as part of routine training without restriction to procedure indication. This study was reviewed by the institutional review board at each participating site and either approved or granted a waiver from the local boards and required only verbal consent from each trainee in accordance with federal regulations [45 C.F.R. 46.117 (c) (2)].[3] Verbal consent was obtained from all participants by their respective site coordinators.

Participants

The participants included all gastroenterology fellows involved in colonoscopy training at each site during the study period. Evaluators were the staff routinely scheduled for teaching on the days when assessment was indicated. All evaluators were credentialed staff GI endoscopists. The site coordinators were tasked with ensuring each participating staff received brief instructions on how to interpret assessment questions and complete the ACE forms.

Assessment tool

The ACE tool consists of 14 questions measuring specific core cognitive and motor skills followed by 2 questions providing an overall assessment of both motor and cognitive skills to act as comparative standards (Table 1). In addition, the form also captures simple demographic data of the procedure (staff and fellow's name) and procedure completion times (time of procedure start, time at which the cecum is reached, and time of procedure completion). The time these events occurred were recorded by the staff (ie, procedure started at 1:30, cecum reached at 1:43, and colonoscope withdrawn at 1:51). From these time marks, cecal intubation times and withdrawal times were calculated (13 and 8 minutes, respectively, from the example).

TABLE 1. Assessment items on the assessment of competency in endoscopy tool

Motor skills
Effective use of air, water, and suction
Scope steering technique
Fine tip control
Loop reduction techniques
Depth of independent scope advancement
Visualization of mucosa
Ability to apply therapeutic tools
Overall motor skills
Cognitive skills
Lumen identification
Knowledge of indication and medical issues
Management of patient discomfort
Pathology identification and interpretation
Identifying location of pathology
Polyp detection
Knowledge of therapeutic tool
Overall cognitive skills

A copy of the colonoscopy ACE assessment tool and detailed description of the scoring metrics can be found in the previously mentioned ASGE white paper.[2] In brief, virtually all of the skills were graded by using a 4-point scale with scores of 1, 2, or 3 (1 = novice, 2 = intermediate, 3 = advanced) demonstrating progression of skills toward, but not yet achieving, minimal competence. A score of 4 (superior) indicated that the fellow demonstrated skills deemed competent to operate independently for that skill during that specific procedure. The 2 exceptions to this were for the metric of depth of independent colonoscope advancement in which a scale of 1 to 8 was used (1 = rectum, 2 = sigmoid, 3 = splenic flexure, 4 = hepatic flexure, 5 = cecum—no terminal ileal attempt, 6 = cecum—failed terminal ileal attempt, 7 = terminal ileum, 8 = other) and independent polyp detection for which a 3-point scale was used (N/A = no polyps present, 1 = none, 2 = some, 3 = all). This metric is used to calculate the polyp detection rate and polyp miss rate for fellows.

Intervention

Before a fellow's participation in the study, the fellow's total number of colonoscopies previously performed was calculated. From that point, each time a fellow reached a multiple of 50 procedures of previous experience (ie, 1, 50, 100,…, 350, 400), the supervising staff was notified to assess the fellow's performance on the next 5 procedures. The staff completed an ACE form at the end of each procedure assessing the fellow's performance. Fellows were not allowed to see the completed graded forms. This was to ensure staff felt free to grade as openly as possible. For all sites, except Mayo Clinic Rochester, assessment was performed on a hard copy version of the form and passed directly back to the site coordinator. The original was then mailed to the principal investigator for inclusion in the database. The Mayo Clinic Rochester site used a Web-based version of the form where, on completion of each form, the data were automatically downloaded directly to a secure encrypted central database. At institutions where the local institutional review board required it, a code was used in place of the names of the fellow and staff. Regardless of the site, all identifiable data of names or institutions were replaced with deidentified labels before analysis.

Analysis

The mean scores of each performance metric were calculated at each of the assessment intervals (every 50th procedure). The differences in the average scores between each stage of training were analyzed by using linear regression predicting motor or cognitive skill scores. An autoregressive generalized estimating equation (with an order of 1) was used to adjust for the institution and to account for within-fellow variability by using a random effect.

For each skill, these average scores define the learning curve in the progression toward competence. The comparisons between stages demonstrate the ACE tool's ability to discriminate between users with differences in experience as small as 50 procedures. Cecal intubation rates were calculated by dividing the total number of cases in which the fellow reached the cecum (score of 5, 6, or 7) by the total number of cases. Depth of insertion scores of 8 (other = indicating postsurgical anatomy or procedures aborted for other reasons) were not included in the calculations of the cecal intubation rate. Independent polyp detection rates were calculated by comparing the number of cases in which the fellow identified at least 1 polyp (score of 2 = some or 3 = all) with the total number of cases. Insertion time, withdrawal time, and total procedure time were also predicted by using regression analysis. Polyp miss rates were calculated by dividing the number of cases in which at least 1 polyp was missed by the fellow (score of 1 = none or 2 = some) by the total number of cases in which polyps were present (score of 1, 2, or 3). Cases in which no polyps were identified by either staff or fellow (defined as N/A = no polyps present) were not included in the denominator of the miss rate calculation. Polyp detection and polyp miss rates were assessed by logistic regression with experience as an independent predictor, and a random effect for fellow regression modeling was performed by using SAS statistical software version 9.3 (SAS Institute Inc, Cary, NC).

To define the minimum thresholds of competency for each skill, the contrasting group standard setting procedure was used.[4] This method defines the score below which users are generally noncompetent and above which users are generally deemed competent. In this procedure, the scores of individual core skills are divided into 2 groups based on whether the fellow was deemed competent (score of 4) or noncompetent (score of 1, 2, or 3) on the corresponding overall motor or cognitive assessment item. For each skill, the distributions of the core skills scores were separately plotted for each group (competent vs noncompetent) by using JMP statistical software version 10.0.0 (SAS Institute Inc) (Fig. 1). The point where the intersection of the competent and noncompetent distribution curves projects onto the score axis (x axis) defines the minimum passing score for the specific metric.

RESULTS

During the academic year July 2013 through June 2014, 184 staff completed 1061 ACE forms grading the colonoscopy skills of 93 gastroenterology fellows at various stages of training (first year, 35%; second year, 22%; third year, 43% of ACE evaluations). Evaluations took place at 10 academic institutions across the United States (Table 2).

Average scores (95% confidence interval [CI]) for each of the 6 motor and 6 cognitive skills graded on the 4-point scale along with the corresponding average overall motor or cognitive competence scores are shown at each assessment interval (Tables 3 and 4, respectively). For each skill, this regression analysis demonstrates measurable and significant improvements in performance scores in relation to experience ($P < .001$). The learning curves for overall motor and cognitive competence are shown in Figure 2.

The thresholds at which minimum competency is first reached, as determined from the contrasting groups method, are reported at the bottom of the respective columns in Tables 3 and 4. These thresholds range from scores of 3.3 to 3.5. Adopting a conservative competency score of 3.5 for all skills allows for uniformity in scoring, ease in interpreting grading yet ensures that the bar meets or exceeds each of the defined thresholds. Additionally, with the bar set at 3.5, the original threshold score identified in the contrasting groups analysis is included in the 95% CI for nearly all skills. This indicates that, by using the corrected 3.5 cutoff, we have 95% confidence that all users have met or surpassed the uncorrected threshold score. The exceptions to this are the ability to apply the tool and accurate location of a lesion where this occurs 1 assessment interval later, and fine tip control where it would occur 1 interval sooner. These thresholds are met at roughly 250 procedures for nearly all skill levels.

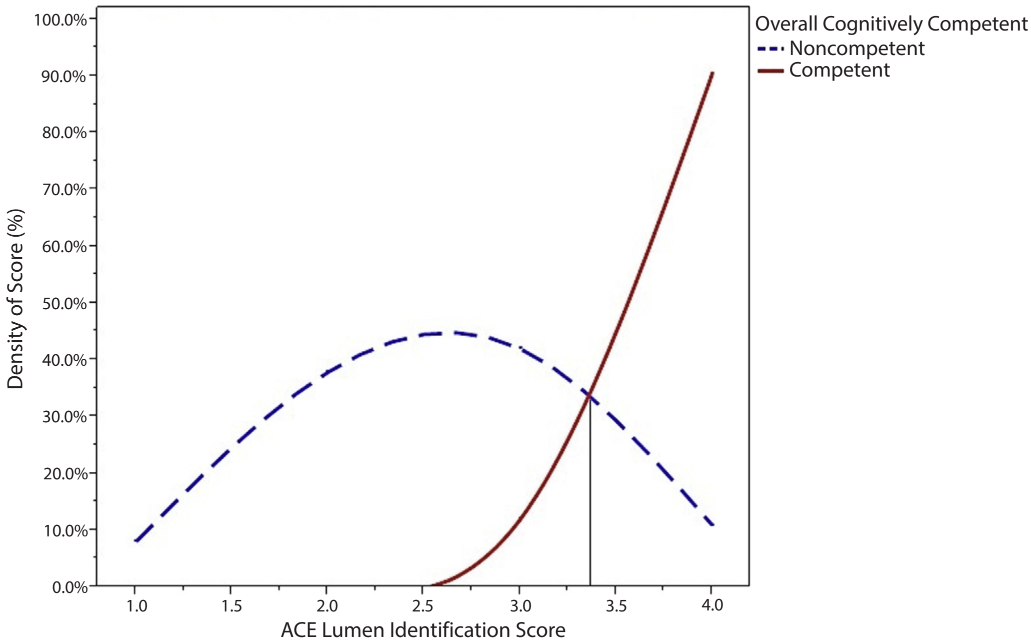

Figure 1. Contrasting groups example. In this standard setting method, a fellow's performance score on a specific skill is placed into either a competent or noncompetent group based on his or her score on the corresponding motor or cognitive overall competence. The distribution curves are plotted of the scores received by each group for each of the assessment of competency in endoscopy (ACE) assessment items. The score (on the *x* axis) that corresponds to where the competent and noncompetent curves intersect defines the minimal passing score for that assessment item. In this example, the distribution of lumen identification scores are plotted for the overall cognitively competent and noncompetent groups.

Independent cecal intubation rates begin at only 17% during initial training and gradually increase to >98% by 400 procedures (Fig. 3). The contrasting groups' analysis identifies that an average independent cecal intubation rate of 90% is needed for the fellow to be deemed competent, confirming the standards broadly accepted in practice. This 90% threshold is achieved at 250 procedures and is shown by the dashed line in Figure 3.

The average time to reach the cecum for beginners is just more than 23.5 minutes, with a total procedure time of 45 minutes (Fig. 4). By 350 procedures, intubation times are less than 10 minutes with a total procedure time less than 30 minutes. By using the contrasting group method, the intersection of the curves (ie, competency thresholds) for these 2 metrics occurred at an intubation time of 15 minutes and a total procedure time of 37 minutes on average. Compared with all of the other metrics, competency in this procedure time occurs somewhat earlier in training at 150 procedures of experience.

Finally, polyp detection and miss rates are shown in Figure 5. In early training, staff identify at least 1 polyp missed by fellows in nearly 80% of cases, but this gradually decreases to near zero toward the end of training. As expected, this is accompanied by a gradual increase in overall polyp detection rates starting at 24% and increasing to 65% by the end of training. The contrasting groups' analysis shows that staff tend to deem fellows competent at the same time that their polyp detection rates reach 50% and the miss rate decreases to less than 25%. Over the entire analysis period, for every additional 50 procedures of experience, there is an improvement in the polyp detection odds ratio of 2.56 (95% CI, 1.55–4.22; $P = .0002$) and the polyp miss rate odds ratio of 0.56 (95% CI, 0.48-0.65; $P < .0001$). This means that for every additional 50 procedures experience, the odds that the fellow would detect at least 1 polyp increased 156% and the odds of missing a polyp decreased by 44%. This is based on the best fit over the entire spectrum of experience (for all procedures from 1 to 450+) because the model used does not allow for analysis over individual intervals.

TABLE 2. Participating institutions

Vanderbilt University
New York University
University of California, San Diego, San Diego, Calif
Scripps Clinic, La Jolla, Calif
Mayo Clinic, Rochester Minn
Mayo Clinic, Jacksonville, Fla
Mayo Clinic, Scottsdale, Ariz
Cleveland Clinic, Cleveland, Ohio
Washington University, St. Louis, Mo
Emory University, Atlanta, Ga

TABLE 3. Motor skills

No. of procedures	N	Use of air, water, and suction	Scope steering technique	Fine tip control	Loop reduction	Mucosal visualization	Ability to apply tool	Overall hands-on competence
0	113	1.6 (1.4-1.7)	1.6 (1.5-1.7)	1.5 (1.4-1.7)	1.4 (1.3-1.5)	1.8 (1.7-2.0)	1.8 (1.5-2.1)	1.4 (1.3-1.5)
50	131	2.4 (2.3-2.5)	2.3 (2.2-2.4)	2.3 (2.1-2.4)	2.3 (2.2-2.4)	2.6 (2.5-2.8)	2.5 (2.4-2.7)	2.2 (2.1-2.3)
100	156	2.9 (2.7-3.0)	2.8 (2.7-2.9)	2.8 (2.7-2.9)	2.8 (2.7-2.9)	3.1 (3.0-3.2)	3.0 (2.8-3.1)	2.7 (2.6-2.8)
150	109	3.2 (3.1-3.3)	3.0 (2.9-3.2)	3.0 (2.9-3.2)	3.1 (3.0-3.2)	3.3 (3.2-3.4)	3.2 (3.0-3.4)	3.0 (2.9-3.1)
200	101	3.3 (3.2-3.5)	3.2 (3.1-3.4)	3.1 (3.0-3.3)	3.2 (3.0-3.3)	3.4 (3.3-3.5)	**3.5 (3.3-3.7)***	3.2 (3.0-3.3)
250	94	**3.5 (3.4-3.7)**	**3.5 (3.4-3.6)**	3.4 (3.3-3.6)†	**3.5 (3.3-3.6)**	**3.7 (3.6-3.8)**	**3.5 (3.4-3.7)**	**3.5 (3.3-3.6)**
300	117	**3.7 (3.6-3.8)**	**3.7 (3.6-3.8)**	**3.6 (3.5-3.7)**	**3.7 (3.6-3.8)**	**3.8 (3.7-3.9)**	**3.7 (3.6-3.8)**	**3.7 (3.6-3.8)**
350	95	**3.7 (3.6-3.8)**	**3.8 (3.7-3.9)**	**3.7 (3.6-3.8)**	**3.8 (3.7-3.9)**	**3.8 (3.7-3.9)**	**3.8 (3.7-3.9)**	**3.8 (3.7-3.9)**
400	70	**3.8 (3.7-3.9)**	**3.8 (3.7-3.9)**	**3.8 (3.7-3.9)**	**3.8 (3.7-3.9)**	**3.9 (3.8-4.0)**	**3.8 (3.7-3.9)**	**3.8 (3.7-3.9)**
450+	75	**3.8 (3.8-3.9)**	**3.8 (3.7-3.9)**	**3.8 (3.7-3.9)**	**3.8 (3.8-3.9)**	**3.9 (3.8-4.0)**	**3.9 (3.8-3.9)**	**3.9 (3.9-4.0)**
P Value for experience		< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
Competency threshold		3.4	3.3	3.3	3.3	3.4	3.4	3.5

The mean scores for the motor skills are shown with the 95% confidence interval at each stage of training along competency threshold calculated with the contrasting groups method.
Bold text indicates the mean score reaches or exceeds the adopted threshold of ≥3.5. Competency is achieved in all motor skills (except for fine tip control) by 250 procedures.
*Indicates a data point where, despite the mean score reaching the adopted 3.5 conservative threshold, the calculated competency threshold (bottom row) still exceeds the lower end of the 95% confidence interval at this point in training.
†Indicates a situation in which the lower end of the 95% confidence interval meets or exceeds the calculated competency threshold despite the mean score not yet reaching the adopted 3.5 conservative threshold.

TABLE 4. Cognitive skills

Experience, no. of procedures	N	Indication issues	Pain management	Landmark identification	Pathology identification	Accurate location of pathology	Knowledge of tool	Overall cognitive competence
0	113	2.9 (2.7-3.0)	2.2 (2.0-2.3)	1.5 (1.4-1.7)	1.9 (1.7-2.1)	2.0 (1.8-2.2)	1.8 (1.4-2.1)	1.8 (1.7-1.9)
50	131	3.3 (3.2-3.4)	2.7 (2.6-2.9)	2.4 (2.3-2.6)	2.7 (2.5-2.8)	2.8 (2.6-2.9)	2.5 (2.3-2.6)	2.5 (2.4-2.6)
100	156	**3.5 (3.5-3.6)**	3.1 (3.0-3.2)	2.9 (2.8-3.0)	3.1 (3.0-3.2)	3.1 (3.0-3.2)	3.0 (2.9-3.2)	3.0 (2.9-3.1)
150	109	**3.6 (3.5-3.7)**	3.3 (3.2-3.4)	3.2 (3.1-3.3)	3.4 (3.2-3.5)	3.3 (3.2-3.4)	3.4 (3.2-3.5)	3.2 (3.0-3.3)
200	101	**3.6 (3.5-3.8)**	3.3 (3.2-3.5)	3.4 (3.2-3.5)	3.4 (3.3-3.5)	**3.5 (3.4-3.6)***	**3.6 (3.4-3.8)**	3.4 (3.3-3.5)
250	94	**3.9 (3.8-4.0)**	**3.5 (3.4-3.7)**	**3.6 (3.5-3.7)**	**3.6 (3.5-3.7)**	**3.7 (3.6-3.9)**	**3.7 (3.5-3.9)**	**3.6 (3.5-3.7)**
300	117	**3.9 (3.9-4.0)**	**3.7 (3.6-3.8)**	**3.8 (3.7-3.8)**	**3.8 (3.7-3.9)**	**3.8 (3.7-3.9)**	**3.7 (3.6-3.9)**	**3.8 (3.7-3.8)**
350	95	**3.9 (3.9-4.0)**	**3.8 (3.7-3.9)**	**3.8 (3.7-3.9)**	**3.8 (3.7-3.9)**	**3.8 (3.7-3.9)**	**3.8 (3.7-3.9)**	**3.8 (3.7-3.9)**
400	70	**3.9 (3.9-4.0)**	**3.8 (3.7-3.9)**	**3.8 (3.7-3.9)**	**3.8 (3.7-3.9)**	**3.9 (3.8-4.0)**	**3.9 (3.8-4.0)**	**3.9 (3.8-4.0)**
450	75	**3.9 (3.8-4.0)**	**3.8 (3.8-3.9)**	**3.8 (3.8-3.9)**	**3.8 (3.7-3.9)**	**3.8 (3.7-3.9)**	**3.9 (3.8-4.0)**	**3.9 (3.9-4.0)**
P value for experience		< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
Competency threshold		3.4	3.4	3.4	3.4	3.5	3.4	3.5

The mean scores for the motor skills are shown with the 95% confidence interval at each stage of training along competency threshold calculated with the contrasting groups method.
Bold text indicates that the mean score reaches or exceeds the adopted threshold of ≥3.5. This is achieved for all cognitive skills by 250 procedures.
*Identifies a data point where, despite the mean score reaching the adopted 3.5 conservative threshold, the calculated competency threshold (bottom row) still exceeds the lower end of the 95% confidence interval at this point in training.

DISCUSSION

In the summer of 2014, the Accreditation Council for Graduate Medical Education put into operation the Next Accreditation System (NAS) for subspecialty training.[5] This reporting system requires programs to monitor and measure trainees' performance throughout training with the goal of establishing competency-based education in medical training. For fellowship programs, this shift requires undertaking the difficult task of defining assessment methods, expected learning curves, and defensible competency endpoints for the skills inherent to each specialty.[6] Training directors have little guidance on how to satisfy these new reporting requirements. However, significant strides have been achieved for colonoscopy training assessment. In this study, we lay out the validity evidence of the new ACE tool (a derivative of the previously reported MCSAT) and describe more

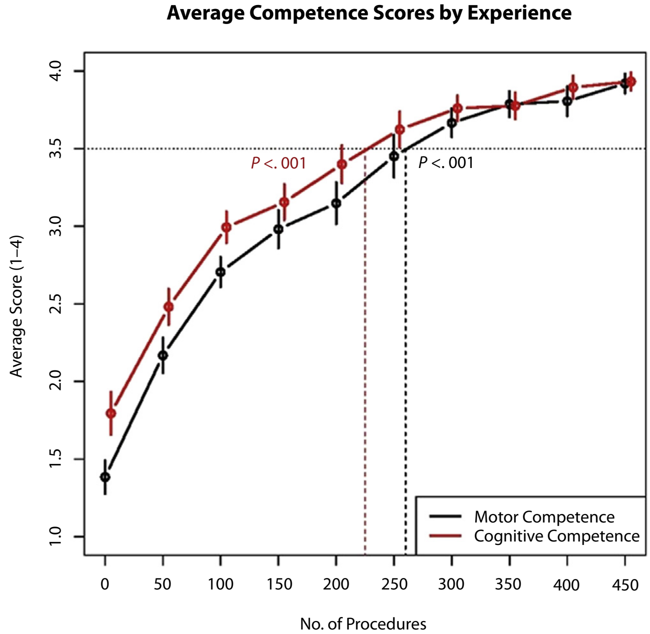

Figure 2. Overall competency scores. The scores from the 2 anchor questions assessing overall motor and cognitive competence are shown (95% confidence interval). An average score of >3.5 suggests the majority of responses were graded as 4 (competent). Although competence in cognitive skills appear to be achieved somewhat sooner than motor skills, both overall scores occur at about 250 procedures coinciding with the point at which competence is achieved in the individual skills as well.

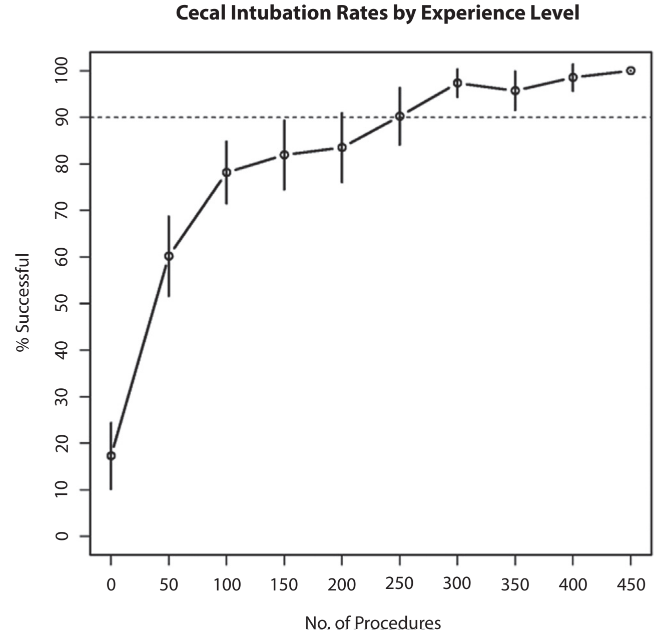

Figure 3. Cecal intubation rates. The cecal intubation rates are strongly associated with the number of procedures performed ($P < .0001$) A contrasting groups analysis identifies an independent cecal intubation rate of 90% as the minimal competence threshold (*dashed line*), matching the value most commonly cited but before this, based on expert opinion alone. Like cognitive and motor skills, this is achieved for the average fellow at 250 procedures.

generalizable competency benchmarks from multiple centers across the country.

In examining the validity evidence of the ACE tool, the constructs for validity evidence set forth by the American Education Research Association are used.[7] Much of the content and response process validity previously reported for the MCSAT tool can be applied to the ACE tool because these use the identical questions and format and are based on the same blueprint of questions as well as their mapping to the desired motor and cognitive skill domains.[8] The new questions of tip control and independent polyp identification found on the ACE tool were developed and vetted by the ASGE Training Committee and were also mapped to one of the previously described skill domains and use the similar scoring systems and descriptors.[2] The relationship to other variables' validity is evidenced by the regression analysis showing the ACE tool's ability to detect a measurable difference in these scores based on differences in experience in as few as 50 procedures. This is a key component to showing that the use of this tool can meaningfully monitor fellows' progress throughout the entire training continuum. This also allows for the establishment of the described generalizable learning curves with which an individual's performance can be compared at any point in training. The obvious advantages of this would be to allow early identification and remediation of fellows falling below the normal learning curves and also to identify specifically which core skills need focused attention. Finally, consequence validity indicates that the results of an assessment have an impact such as passing or failing a rotation or being credentialed. At the Mayo Clinic Rochester, all fellows are expected to meet or exceed these competency performance benchmarks to graduate fellowship.

The ACE competency benchmarks are summarized in Table 5. Despite that these were obtained from a larger and more diverse cohort of fellows than the original MCSAT report, the results are very similar. As before, fellows are expected to achieve average scores of $\geq$3.5 for each core skill. Average cecal intubation times decreased slightly from less than 16 minutes to now less than 15 minutes. The independent cecal intubation rate increased modestly from 85% in the original report to 90% with this larger cohort. This ACE threshold matches the 90% generally cited by experts and training documents but until now was primarily based on expert opinion.[9,10] These new benchmarks are obtained by the average fellow at approximately 250 procedures experience and by 90% of fellows by 300 procedures. This is not to suggest these procedure volumes are to be used to determine competency but rather simply as a guide in curriculum planning to ensure that fellows receive enough endoscopy time to achieve the desired benchmarks.

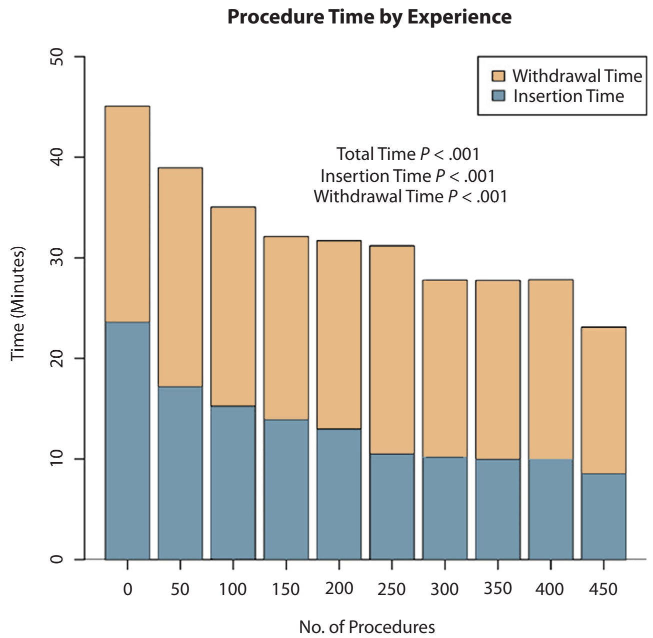

Figure 4. Procedure times. The mean insertion, withdrawal, and total procedure times are shown in this stacked bar graph. Cecal intubation times of less than 15 minutes are defined by the standard setting methods. This is typically achieved at about 100 procedures, but the times continue to improve down to a plateau of about 9 minutes. Withdrawal times do not appear to vary much.

Figure 5. Polyp detection rates. The polyp detection and miss rates are shown (95% confidence interval). As expected, there are high miss rates early on in training. By the end of training, miss rates decrease greatly (2%), along with a concomitant increase in polyp detection rates (65%).

The ACE tool introduces polyp detection metrics not present in the original MCSAT form. The previous omission was for 2 reasons. At the time of the MCSAT study, polyp detection rates were not thought to be an adequate reflection of adenoma detection rates and determining that pathology results was simply too burdensome. Although still not ideal, polyp detection rates have been subsequently shown to be useful as a possible surrogate for adenoma detection rates. Polyp miss rates reported in this study, however, may still underreport true miss rates due to the fact that the fellow alone managed the withdrawal of the colonoscope, thereby possibly obscuring polyps to even the supervising staff.

The scores and learning curves established here are useful for the completion of the NAS fellow reports (specifically for the metric 4a, demonstrates skill in performing and interpreting invasive procedures). In the NAS reporting system, trainees are graded on a continuum from critical deficiencies to ready for unsupervised practice with descriptors for each of these first 4 categories. As with the ACE tool, scoring in the NAS is not intended to be a norm-referenced assessment in which a trainee's performance is graded in relation to one's peers. Instead, the emphasis is on competency-based education outcomes in which performance is graded on a criterion-referenced spectrum of skills from novice to competent. In essence, beginning trainees are expected to receive scores in the critical deficiencies range at the very start of training and gradually increase to scores of ready for unsupervised practice by the end of training. The 4 performance levels and their descriptors found on the ACE and MCSAT tools correlate with the first 4 performance categories of the NAS continuum. As such, ACE scores could be used directly to complete aspects of the NAS reports. One caveat to this is that on the NAS report, there is a fifth aspirational category for the rare instance in which a trainee meets or exceeds the skills expected of junior staff. The ACE tool is not designed to assess skills beyond those of simply meeting minimal competence. As such, determining whether a competent trainee's skills are aspirational would still require the subjective assessment by the program director and staff.

How a program implements the ACE tool may differ. Whether assessments are completed for every procedure or only at set intervals, as was done in this study, or whether hard copy forms, online forms, or integration into endoscopy software is used will depend on specific programs' resources and abilities. As we have shown, interval assessments such as every 50 procedures are certainly a viable option when continuous assessment is not possible. This method, however, requires fellows' procedure numbers to be continuously monitored to determine when assessment is indicated.

The strengths of this study rest in the size and diversity of the study group. With nearly 100 fellows being drawn from 10 different centers of varying sizes and located fairly evenly across the country, the results should be fairly generalizable to most gastroenterology fellows training in

TABLE 5. Summary of competency benchmarks

Independent cecal intubation rate	≥90%
Cecal intubation time	≤15 min
Each ACE skill, average score	≥3.5
Polyp detection rate	≥50%
Polyp miss rate	≤25%

ACE, Assessment of competency in endoscopy.

U.S. centers. When interpreting these results, however, one needs to remain cognizant of the study's weaknesses, particularly the fact that teaching staff were not blinded to the fellow's level of training or degree of previous experience. Accomplishing this would require a group of staff from outside institutions (ie, unfamiliar with the trainee being assessed) being provided the time and resources to observe and assess procedures either in person or recorded interactions. Apart from the obvious amount of time and resources that this would take to assess over 1000 procedures, there is also the issue of patient confidentiality and consent to be recorded or observed. All of these issues would make a blinded study nearly impossible to design and fund.

Another limitation of this study is that it focused only on gastroenterology fellows; hence, extrapolation of these learning curves to general surgery residents, colorectal surgery fellows, or family practice physicians who also train in colonoscopy is not possible without further research. Such research would need to control for potential differences in the interpretation of acceptable skills by staff of varying experience. To truly compare fellows of different training pathways (gastroenterology, surgical, family practice), the assessing staff would have to be the same for all groups, such as institutions where the gastroenterology staff train the fellows of the various groups. Alternatively, having multiple staff from multiple departments assess each procedure would determine whether indeed there are differences between the fellow's performance based on training pathway and whether there are differences in staff grading based on specialty. This type of study would go far in helping to develop a unified set of competency metrics by which all training pathways would abide.

This study reports, and provides validity evidence of, the learning curves and competency benchmarks for colonoscopy training in gastroenterology fellows as determined by a large and diverse group of trainees. The results closely mirror the previous MCSAT study results and support the quality metrics for colonoscopy derived from expert opinion and laid out in current gastroenterology training guidelines. The next step is to determine whether these learning curves and metrics are applicable to trainees of alternate training pathways and to derive a unified set of competency endpoints that all those seeking privileges in colonoscopy must meet regardless of specialty.

REFERENCES

1. Sedlack RE. Training to competency in colonoscopy: assessing and defining competency standards. Gastrointest Endosc 2011;74:355-66.
2. ASGE Training Committee. ASGE's assessment of competency in endoscopy evaluation tools for colonoscopy and EGD. Gastrointest Endosc 2014;79:1-7.
3. US Department of Health and Human Services. Code of Federal Regulations. Title 45, Part 46. Section 117. Available at: http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html#46.117. Accessed December 23, 2014.
4. Downing SM, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. Teach Learn Med 2006;18:50-7.
5. Accreditation Council for Graduate Medical Education; American Board of Internal Medicine. The Internal Medicine Subspecialty Milestone Project. Available at: https://www.acgme.org/acgmeweb/Portals/0/PDFs/Milestones/InternalMedicineSubspecialtyMilestones.pdf. Accessed January 28, 2015.
6. Northrup PG, Argo CK, Muir AJ, et al. Procedural competency of gastroenterology trainees: from apprenticeship to milestones. Gastroenterology 2013;144:677-80.
7. American Education Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington (DC): American Education Research Association; 2014.
8. Sedlack RE. The Mayo Colonoscopy Skills Assessment Tool: validation of a unique instrument to assess colonoscopy skills in trainees. Gastrointest Endosc 2010;72:1125-33.
9. Rex DK, Schoenfeld PS, Cohen J, et al. Quality indicators for colonoscopy. Am J Gastroenterol 2015;110:72-90.
10. ASGE Training Committee. Principles in training in GI endoscopy. Gastrointest Endosc 2012;75:231-5.

APPENDIX A. ACE RESEARCH GROUP

Robert E. Sedlack, MD, MHPE, Mayo Clinic, Rochester, Minnesota

Walter J. Coyle, MD, Scripps Clinic, La Jolla, California

Keith L. Obstein, MD, MPH, Vanderbilt University Medical Center, Nashville, Tennessee

Michael A. Poles, MD, PhD, New York University School of Medicine, New York, New York

Francisco C. Ramirez, MD, Mayo Clinic, Scottsdale, Arizona

Frank J. Lukens, MD, Mayo Clinic, Jacksonville, Florida

C. Prakash Gyawali, MD, Washington University, St. Louis, Missouri

Jennifer A. Christie, MD, Emory University, Atlanta, Georgia

Denise Kalmaz, MD, University of California San Diego, San Diego, California

Carol A. Burke, MD, Cleveland Clinic, Cleveland, Ohio

Felicity Enders, PhD (Biostatistics), Mayo Clinic, Rochester, Minnesota

Joseph J. Larson, BS (Biostatistics), Mayo Clinic, Rochester, Minnesota

Amy S. Oxentenko, MD, Mayo Clinic, Rochester, Minnesota