The Gastrointestinal Endoscopy Competency Assessment Tool for Pediatric Colonoscopy

*Catharine M. Walsh, *Simon C. Ling, †Petar Mamula, ‡Jenifer R. Lightdale, *Thomas D. Walters, §Jeffrey J. Yu, and ||Heather Carnahan

See ''GiECAT$_{KIDS}$ Validated Pediatric Colonoscopy Assessment Tool: A Call to Action'' by Sauer and Narkewicz on page 425.

ABSTRACT

Objectives: Validated assessment tools are required to support competency-based education. We aimed to assess the reliability and validity of the Gastrointestinal Endoscopy Competency Assessment Tool for Pediatric Colonoscopy (GiECAT$_{KIDS}$), an instrument developed by 41 North American experts using Delphi methodology.

Methods: GiECAT$_{KIDS}$ consists of a 7-item global rating scale (GRS) and an 18-item checklist (CL). An attending physician assessed 104 colonoscopies performed at 3 North American hospitals by 56 endoscopists, including 25 novices (<50 previous procedures), 21 intermediates (50–250), and 10 advanced endoscopists (>500). Another observer rated procedures to assess interrater reliability using intraclass correlation coefficient (ICC). Test–retest reliability was measured with ICC comparing endoscopists' first and second procedure scores. Discriminative validity was examined by comparing experience level with scores. Concurrent validity was assessed by correlating scores with colonoscopy experience, cecal and terminal ileal intubation rates, and physician global assessment.

Results: Interrater reliability of the GiECAT$_{KIDS}$ was high (total: ICC = 0.88; GRS: ICC = 0.79; CL: ICC = 0.89). Test–retest reliability was excellent (total: ICC = 0.94; GRS: ICC = 0.94; CL: ICC = 0.84). GiECAT$_{KIDS}$ total, GRS, and CL scores differed significantly among novice, intermediate, and advanced endoscopists ($P < 0.001$). There was a significant positive correlation ($P < 0.001$) between scores and number of previous colonoscopies (total: $\rho = 0.91$, GRS: $\rho = 0.92$, CL: $\rho = 0.84$), cecal intubation rate (total: $\rho = 0.82$, GRS: $\rho = 0.85$, CL: $\rho = 0.77$), ileal intubation rate (total: $\rho = 0.82$, GRS: $\rho = 0.82$, CL: $\rho = 0.80$), and physician global assessment (total: $\rho = 0.95$, GRS: $\rho = 0.94$, CL: $\rho = 0.89$).

Conclusions: The GiECAT$_{KIDS}$ demonstrates strong reliability and validity as a measure of performance of pediatric colonoscopy that can be used to support training and assessment.

Key Words: clinical competence, education, medical, graduate/standards, educational measurement, endoscopy, gastrointestinal/education, endoscopy, gastrointestinal/standards, endoscopy, pediatric

(*JPGN* 2015;60: 474–480)

Around the globe, postgraduate medical education is implementing competency-based reforms, shifting training from a time and process-based paradigm to an educational process intended to result in demonstrated training outcomes (1). Competency-based education implies that residents remain in training until they have acquired the requisite core knowledge, skills, and attitudes and can apply them independently (2). Despite the defined movement toward competency-based curricula and outcome evaluation, there remains a paucity of rigorously developed and validated tools across a number of disciplines, including pediatric endoscopy.

Competence in pediatric colonoscopy requires specialized training and practice (3). Delivery of safe and high-quality endoscopic care requires proficiency in 3 main competency domains: technical (psychomotor), cognitive, and integrative competencies (eg, diagnostic reasoning and communication) (4). Integrative competencies are higher-level competencies required to perform a procedure that complement an individual's technical skills and clinical knowledge to facilitate effective delivery of safe and appropriate care in varied contexts. Present North American licensure requirements (5,6) include end-of-training examinations but no formal assessment of endoscopic competence. During training, endoscopic skills are typically evaluated as part of in-training evaluation reports (ITERs) that summarize a trainee's performance during a clinical rotation. ITERs, however, have been shown to suffer from poor reliability and are prone to recall bias and ''halo'' effects, whereby performance in 1 area can bias judgment in other

Received June 4, 2014; accepted December 18, 2014.

From the *Department of Paediatrics, Division of Gastroenterology, Hepatology and Nutrition, The Hospital for Sick Children, University of Toronto, Ontario, Canada, the †Division of Gastroenterology, Hepatology and Nutrition, The Children's Hospital of Philadelphia, PA, the ‡Department of Pediatrics, Division of Pediatric Gastroenterology and Nutrition, UMass Memorial Children's Medical Center, University of Massachusetts, Worcester, MA, the §Wilson Centre, Faculty of Medicine, University of Toronto, Ontario, and the ||School of Human Kinetics and Recreation, Memorial University of Newfoundland, St John's, Newfoundland, Canada.

Address correspondence and reprint requests to Catharine M. Walsh, MD, PhD, Division of Gastroenterology, Hepatology and Nutrition, Hospital for Sick Children, 555 University Ave, Room 8417, Black Wing, Toronto, ON, Canada M5G 1X8 (e-mail: catharine.walsh@mail.utoronto.ca).

Supplemental digital content is available for this article. Direct URL citations appear in the printed text, and links to the digital files are provided in the HTML text of this article on the journal's Web site (*www.jpgn.org*).

The abstract of an earlier version of this article was presented at the 2013 North American Society for Pediatric Gastroenterology, Hepatology, and Nutrition Annual Meeting, the 2014 Ottawa Conference, and the 2014 Digestive Diseases Week conference.

This project was supported by an American Society of Gastrointestinal Endoscopy Quality in Endoscopic Research Award. C.M.W. is a doctoral fellow of the CIHR Canadian Child Health Clinician Scientist Program, the recipient of a Department of Paediatrics Research Fellowship (The Hospital for Sick Children), and a Postgraduate Medical Education Award, University of Toronto. The other authors report no conflicts of interest.

DOI: 10.1097/MPG.0000000000000686

aspects of performance (7). In addition, ITERs provide learners with summative reports, instead of specific information that can be used as feedback to monitor and modify performance and improve learning (8). Log books, used by endoscopists to record their clinical experiences, are another common assessment method; however, the objectivity and accuracy of the records have been questioned (7,9). In addition, case logs reflect procedural volume, which is not necessarily indicative of operative ability because individuals learn at different rates (10).

An ideal tool for the assessment of endoscopic competence should be feasible, reliable, valid, acceptable, and cost-effective and achieve a desirable educational impact (11). Tools to measure clinical ability in performing colonoscopy have been produced, but there is at present no measure of endoscopic competence that has been validated specifically within the pediatric context. Use of the Delphi consensus technique enabled our group to develop the Gastrointestinal Endoscopy Competency Assessment Tool for Pediatric Colonoscopy ($GiECAT_{KIDS}$), a pediatric-specific measure of endoscopic competence that is reflective of practice across North America and was designed to assess procedure-related technical, cognitive, and integrative competencies in a continuous fashion throughout training (12). The objective of the present study is to assess the reliability and validity of the $GiECAT_{KIDS}$ for use in evaluating the competence of clinicians performing colonoscopy in pediatric patients in the clinical setting.

METHODS

The present study was a prospective North American multicenter study designed to assess the reliability and validity of the $GiECAT_{KIDS}$. Ethical approval was obtained from The Hospital for Sick Children's research ethics board, University of Toronto research ethics board, and Boston Children's Hospital's institutional review board. The Children's Hospital of Philadelphia's institutional review board granted ethics exempt status as a quality improvement project. Informed consent was obtained from all of the participants where required.

Participants

Study participants were pediatric gastroenterology fellows and faculty from 3 North American academic teaching hospitals. Purposive sampling was used to recruit participants for a novice, an intermediate, and an advanced group according to prespecified procedure volume criteria, which were based on present credentialing guidelines and a literature review of endoscopic competence (13). Novice endoscopists were defined as individuals who had performed <50 previous colonoscopies. Participants were considered intermediate endoscopists if they had performed between 50 and 250 colonoscopy procedures. Advanced endoscopists were those who had performed >500 colonoscopies.

Performance Measure

The $GiECAT_{KIDS}$ was developed by a panel of 41 pediatric endoscopy experts from 28 centers across North America using Delphi consensus methodology (12). It is composed of a global rating scale (GRS) that assesses more holistic aspects of the skill and a highly structured checklist (CL) that outlines key steps required to complete the procedure (Appendix 1, *http://links.lww.com/MPG/A441*). The GRS rates 7 domains (technical skill, strategies for scope advancement, visualization of mucosa, independent procedure completion [need for assistance], knowledge of procedure, interpretation and management of findings, and patient safety) using a 5-point Likert scale with descriptive anchors reflective of the degree of autonomy demonstrated by the endoscopist (ie, the degree to which the endoscopist required verbal and/or hands-on guidance to complete the task(s)). The sum of scores of each of the 7 items yields a total score from 7 to 35, with higher scores reflecting a better performance. The 18-item CL scores each item on a dichotomous scale (1 = performed correctly or 0 = not performed/performed incorrectly) with total score ranging from 0 to 18, modeled after a previously validated CL scoring system used in general surgery (14).

Data Collection

Each endoscopist was assessed in real time performing 2 colonoscopies, <2 weeks apart, in the clinical setting. Novice and intermediate endoscopists were gastroenterology fellows supervised by an attending endoscopist who provided verbal and/or hands-on guidance and/or took over any part of the procedure as per usual practice. The endoscopist's performance during each colonoscopy was assessed by 1 experienced attending endoscopist using the $GiECAT_{KIDS}$ assessment tool. There were different assessors for different endoscopists. Before using the $GiECAT_{KIDS}$ for the first time, a trained observer reviewed the items with each assessor. Raters were encouraged to read the description of each domain and to use the full range of responses, but no formal rater training was undertaken. Using 5-point Likert scales, the assessor was also asked to provide an overall physician global assessment (PGA) of endoscopic competence and an overall global assessment of the endoscopist's technical, cognitive, and integrative skills. A second trained observer (site leads) rated a subset of procedures independently to determine interrater reliability.

Participants completed a demographic questionnaire including level of training (if applicable), hand dominance, sex, number of years performing colonoscopy, an estimate of their experience in performing colonoscopy, and their independent cecal and terminal ileal intubation rates based on the previous 20 colonoscopies they had performed.

Reliability Analysis

Three measures were used for the analysis of internal structure of the $GiECAT_{KIDS}$: item-total correlations, interitem correlation, and internal consistency (using Cronbach α and Kuder-Richardson formula 20 for the GRS and CL, respectively) (15). Internal structure of the $GiECAT_{KIDS}$ was also analyzed by comparing an endoscopist's total combined score for the technical (GRS items 1–4, 7; CL items 5–11), cognitive (GRS item 5, CL items 1, 3, 5, 12, 13, 15), and integrative (GRS items 6, 7; CL items 1, 2, 4, 14–18) items with their respective overall physician global ratings of technical, cognitive, and integrative skills using Pearson product-moment correlation coefficient. Total CL and GRS scores were also correlated using Pearson correlation coefficient.

Interrater reliability was assessed based on the 2 $GiECAT_{KIDS}$ scores assigned independently by the attending endoscopist and trained observer for a single colonoscopy procedure. Interrater reliability was determined by the intraclass correlation coefficient model 1 ($ICC_{1,1}$) using a 1-way random-effects model for both single measures (individual rater) and average measures (the average of 2 raters' scores) for the total $GiECAT_{KIDS}$, GRS, and CL scores (16,17). This model is suitable when each participant is assessed by a different set of randomly selected raters, and it ensures generalizability of the findings to other raters with similar characteristics (16,17).

Test–retest reliability was established by comparing the $GiECAT_{KIDS}$ scores given for an endoscopist's first and second procedures. Test–retest reliability of the total $GiECAT_{KIDS}$, GRS,

and CL scores was assessed using the intraclass correlation coefficient model 2 ($ICC_{2,1}$) using a 2-way random-effects model (participant by procedure, single measures) and the absolute agreement definition (16,18).

Validity Analysis

Only data from each endoscopist's first procedure were included in the validity analysis to avoid bias because 2 procedures were not captured for all of the participants. Discriminative validity of the $GiECAT_{KIDS}$ was assessed by comparing the total $GiECAT_{KIDS}$, GRS, and CL scores of novice, intermediate, and advanced endoscopists using separate Kruskal-Wallis tests with post hoc Bonferroni-corrected Mann-Whitney *U* pairwise comparisons. In order to further evaluate discriminative validity, receiver operating characteristic (ROC) curves were generated to assess the ability of the $GiECAT_{KIDS}$ to differentiate between endoscopists who were assigned an overall rating of "competent" (PGA score of 4 or 5) and those who were deemed not yet competent (PGA score of 1, 2, or 3).

To examine concurrent validity, total $GiECAT_{KIDS}$, GRS, and CL scores were correlated, using Spearman correlation coefficient, with colonoscopy experience, cecal intubation rates, terminal ileal intubation rates, and PGA of skill. Finally, the total $GiECAT_{KIDS}$, GRS, and CL scores were plotted against procedure numbers.

Statistical Analysis

All of the statistical analyses were performed using SPSS version 20.0 (IBM SPSS Statistics, Armonk, NY). A *P* value of ≤ 0.05 was considered statistically significant. Sample size was determined based on the number of participants required to obtain statistically significant differences in the scores of novice, intermediate, and advanced endoscopists in the analysis of discriminative validity. A power calculation was performed a priori based on work with the objective assessment of technical skills, a surgical direct observation assessment tool (14). Using power of 0.8, an α of 0.05, and an effect size of 0.6, the minimum required number of participants for each group was 10; therefore, recruitment continued until data had been collected on at least 10 endoscopists per group.

RESULTS

Data were collected on 104 colonoscopies performed at 3 North American teaching hospitals by 56 endoscopists, comprising 25 novices, 21 intermediates, and 10 advanced endoscopists. The characteristics of the cohort are displayed in Table 1. Novice endoscopists had performed on average 15.3 ± 13.5 ($\pm$ standard deviation) previous colonoscopies and had a self-reported mean cecal intubation rate of $12.6\% \pm 21.12\%$ and a mean terminal ileal intubation rate of $4.56\% \pm 7.96\%$. Intermediate endoscopists had performed on average 116.1 ± 44.3 previous colonoscopies and had a mean cecal intubation rate of $77.7\% \pm 21.3\%$ and a mean terminal ileal intubation rate of $66.3\% \pm 28.3\%$. Advanced endoscopists had performed on average 607.8 ± 191.7 previous colonoscopies and had a mean cecal intubation rate of $97.0\% \pm 3.1\%$ and a mean terminal ileal intubation rate of $93.6\% \pm 3.4\%$.

Reliability Analysis

Internal consistency was 0.98 for the $GiECAT_{KIDS}$ GRS, with α values ranging from 0.97 to 0.98 when each item was deleted. Interitem correlations for the GRS ranged from 0.77 to 0.92, and item-total correlations ranged from 0.87 to 0.95. Item discrimination statistics for the GRS are provided in Table 2 along with the correlation between each item and overall PGA of skill. Internal consistency was 0.87 for the $GiECAT_{KIDS}$ CL and was also found to remain consistent when tested against deleting each CL item (α values 0.85–0.87). There was a high degree of correlation between the total combined technical, cognitive, and integrative item scores and the overall physician global ratings of technical, cognitive, and integrative skills with correlation values of 0.94, 0.85, and 0.91, respectively ($P < 0.001$). A significant positive correlation was found between total GRS and CL scores ($r = 0.88$, $P < 0.001$).

Attending endoscopist's and observer's evaluations were obtained for 22 colonoscopies (11 novice and 11 intermediate endoscopists). The interrater reliability for the attending and observer was high for the $GiECAT_{KIDS}$ total, GRS, and CL scores (Table 3). A total of 48 of the 56 endoscopists were assessed performing 2 colonoscopies (21 novices, 20 intermediates, and 7 advanced endoscopists). The mean time between an endoscopist's first and second procedures was 3.70 ± 4.98 days. The test–retest reliability coefficients for the $GiECAT_{KIDS}$ were excellent for total, the GRS, and CL scores (Table 4).

Validity Analysis

Analysis of total $GiECAT_{KIDS}$ scale scores showed a significant main effect of group (novice, intermediate, advanced; Kruskal-Wallis $= 42.35_2$, $P < 0.001$, $\eta^2 = 0.77$; Table 5). Post hoc analysis revealed that total $GiECAT_{KIDS}$ scale scores differed significantly between each group ($P < 0.001$). GRS scores also increased with the level of expertise (Kruskal-Wallis $= 43.74_2$, $P < 0.001$, $\eta^2 = 0.80$; Table 5). Post hoc analysis showed that the advanced group scored significantly higher than the intermediate group ($P = 0.001$), which scored higher than the novice group ($P < 0.001$). Finally, there was also a significant main effect of

TABLE 1. Endoscopist participant characteristics

	Demographic characteristic									
	Training level, % (n)		Hand dominance, % (n)		Sex, % (n)		Number of years performing colonoscopy, % (n)			
	GI fellow	GI attending	Right	Left	Male	Female	<1	1–5	6–10	>10
Overall	82.1 (46)	17.9 (10)	91.1 (51)	8.9 (5)	41.1 (23)	58.9 (33)	46.4 (26)	35.7 (20)	7.1 (4)	10.7 (6)
Novice (25)	100.0 (25)	0	96.0 (24)	4.0 (1)	32.0 (8)	68.0 (17)	100.0 (25)	0	0	0
Intermediate (21)	100.0 (21)	0	90.5 (19)	9.5 (2)	42.9 (9)	57.1 (12)	4.8 (1)	5.2 (20)	0	0
Advanced (10)	0	100.0 (10)	80.0 (8)	20.0 (2)	60.0 (6)	40.0 (4)	0	0	40.0 (4)	60.0 (6)

GI = gastroenterology.

TABLE 2. Interitem and item-total correlations for the GiECAT$_{KIDS}$ GRS

GRS dimension	TS	SA	VM	PC	K	IMF	Item-total correlation	Correlation with physician global assessment of skill
TS							0.94	0.92
SA	0.92						0.94	0.89
VM	0.91	0.89					0.91	0.90
PC	0.88	0.87	0.86				0.92	0.92
K	0.83	0.85	0.77	0.82			0.87	0.91
IMF	0.90	0.91	0.85	0.92	0.89		0.95	0.84
PS	0.86	0.86	0.83	0.87	0.80	0.89	0.90	0.86

GiECAT$_{KIDS}$ = Gastrointestinal Endoscopy Competency Assessment Tool for Pediatric Colonoscopy; GRS = global rating scale; IMF = interpretation and management of findings; K = knowledge of procedure; PC = independent procedure completion; PS = patient safety; SA = strategies for endoscope advancement; TS = technical skill; VM = visualization of mucosa.

group for CL scale scores (Kruskal-Wallis = 35.13_2, $P < 0.001$, $\eta^2 = 0.64$; Table 5). Post hoc planned comparisons also indicated that scores were significantly higher among the advanced group as compared with those among the intermediate group ($P = 0.011$), which scored higher than the novice group ($P < 0.001$).

When we compared endoscopists with ''competent'' PGA scores (4 or 5) versus ''noncompetent'' PGA scores (1, 2, or 3), the areas under the ROC curve for total GiECAT$_{KIDS}$, GRS, and CL scores were 0.99 (95% confidence interval [CI], 0.96–1.00), 0.98 (95% CI, 0.95–1.00), and 0.99 (95% CI, 0.97–1.00), respectively. High areas under the ROC curve indicate strong discriminatory performance of the GiECAT$_{KIDS}$ assessment tool.

There was a significant positive correlation ($P < 0.001$) between GiECAT$_{KIDS}$ scores and number of previous colonoscopies, cecal intubation rate, ileal intubation rate, and PGA of skill (Table 6).

As can be seen from Figure 1, the scatterplot of total GiECAT$_{KIDS}$ scores versus number of previous colonoscopies performed follows the shape of a typical learning curve with rapidly improving performance early in the learning process and flattening out to an asymptote because the rate of improvement slows with experience. Plots of procedure numbers versus GiECAT$_{KIDS}$ GRS and CL scores revealed a similar shape (graphs not shown).

DISCUSSION

The present study establishes the reliability and validity of the GiECAT$_{KIDS}$, a pediatric-specific assessment tool for colonoscopy, in the context of live procedures. Reliability refers to the consistency or reproducibility of a set of measurements. Interrater reliability reflects the ability of an assessment tool to produce consistent results when an individual is rated by multiple independent assessors at the same time (19). Although the GiECAT$_{KIDS}$ demonstrated high interrater reliability with all of the values meeting the cutoff of 0.75 that is generally considered to indicate good reliability in educational measurement (20), interrater reliability for a single rater did not exceed 0.9, which is traditionally considered acceptable high reliability for very-high-stakes assessments such as medical credentialing or licensure examinations that have major consequences for both the examinees and the society (21). Unreliability of measurement tools generally arises from 3 sources: the patient, the procedure, and the rater (22). In the present study, the use of different raters for different endoscopists likely reduced interrater reliability as compared with a study design in which the same 2 raters assessed all of the endoscopists. In addition, the attending endoscopists who performed one of the ratings were deliberately not calibrated or extensively trained to reflect conditions normally occurring during training. Rater training and rater practice would likely increase interrater reliability of the GiECAT$_{KIDS}$, because both have been shown to be effective strategies to improve reliability (22). Furthermore, interrater reliability was calculated based on a subset of 22 procedures performed by 11 novice and 11 intermediate endoscopists, a factor that may have affected reliability. Test–retest reliability represents consistency in an assessment by the same rater (single measures ICC) or a set of raters (average measures ICC) over time (16). The test–retest reliability of the GiECAT$_{KIDS}$ was also high, well above the acceptable threshold of 0.8 (18,23).

The high internal consistency, and interitem and item-total correlations of the GiECAT$_{KIDS}$ GRS suggest that the 7 items measure a single construct of ''endoscopic competence''; yet the scale assesses diverse dimensions such as technical skill,

TABLE 3. Interrater reliability coefficients for the total GiECAT$_{KIDS}$ score, global rating scale, and checklist components

Component of GiECAT$_{KIDS}$ scale	ICC$_{1,1}$, single measures		ICC$_{1,1}$, average measures (2 raters)	
	ICC	95% CI	ICC	95% CI
Total GiECAT$_{KIDS}$ score*	0.88	0.74–0.95	0.94	0.85–097
Global rating scale score*	0.79	0.56–0.91	0.88	0.72–0.95
Checklist score*	0.89	075–0.95	0.94	0.86–0.98

CI = confidence interval; GiECAT$_{KIDS}$ = Gastrointestinal Endoscopy Competency Assessment Tool for Pediatric Colonoscopy; ICC = intraclass correlation coefficient.
* Correlations significant ($P < 0.001$).

TABLE 4. Test–retest reliability coefficients for the total GiECAT$_{KIDS}$ score, global rating scale, and checklist components

Component of GiECAT$_{KIDS}$ scale	ICC$_{2,1}$, single measures		ICC$_{2,1}$, average measures (2 raters)	
	ICC	95% CI	ICC	95% CI
Total GiECAT$_{KIDS}$ score*	0.94	0.90–0.97	0.97	0.95–0.98
Global rating scale score*	0.94	0.90–0.97	0.97	0.95–0.98
Checklist score*	0.84	0.74–0.91	0.92	0.85–0.95

CI = confidence interval; GiECAT$_{KIDS}$ = Gastrointestinal Endoscopy Competency Assessment Tool for Pediatric Colonoscopy; ICC = intraclass correlation coefficient.
* Correlations significant ($P < 0.001$).

TABLE 5. Comparison of scores between groups for total $GiECAT_{KIDS}$, global rating scale, and checklist components using procedure volume criteria for definition of novice, intermediate, and advanced endoscopists

Component of $GiECAT_{KIDS}$ scale	Score			P*	Maximum possible score
	Novice	Intermediate	Advanced		
Total $GiECAT_{KIDS}$ score†	22.00 (10.50)	44.00 (7.00)	51.00 (2.25)	<0.001	53
Global rating scale score†	14.00 (7.00)	27.00 (5.50)	34.00 (1.00)	<0.001	35
Checklist score†	9.00 (4.00)	16.00 (2.50)	17.00 (2.00)	<0.001	18

Scores reported as medians (interquartile range). $GiECAT_{KIDS}$ = Gastrointestinal Endoscopy Competency Assessment Tool for Pediatric Colonoscopy.
* Significant differences between groups ($P < 0.001$). Comparisons were carried out using Kruskal-Wallis test of ranks.
† Post hoc Mann-Whitney U pairwise comparisons showed significant differences among novice, intermediate, and advanced endoscopists ($P < 0.02$).

knowledge, and clinical judgment (19). This is in line with previous research that suggests that assessors have difficulty distinguishing between dimensions of competence (24). Psychologically, raters may be forming a coherent impression of a learner based on the information they are receiving that influences subsequent clinical judgments (25). The high degree of correlation between the combined technical, cognitive, and integrative global rating and CL items with their respective scores of overall technical, cognitive, and integrative skills, however, suggests that these indicators act as reliable measures of each respective competency domain (technical, cognitive, integrative), independent of the overall assessment.

Construct validity, a core property of a measurement tool, refers to the ability of an assessment method to measure the concept it is intended to measure (26). Construct validity was examined using both discriminative validity (which reflects whether the measure differentiates differences in skill across groups hypothesized to score differently) and concurrent validity (which reflects the relation of the index with other variables measuring a similar attribute) (26). Evidence of discriminative validity was demonstrated by statistically significant differences in scores of novice, intermediate, and advanced endoscopists and a strong ability of the $GiECAT_{KIDS}$ tool to discriminate endoscopists who were assigned a PGA rating of ''competent'' against those who were deemed not yet competent. The high correlation of $GiECAT_{KIDS}$ scores with other measures of ''endoscopic competence,'' including self-reported colonoscopy experience, cecal and terminal ileal intubation rates, and PGA of skill, provides psychometric evidence of concurrent validity. The advantage of the $GiECAT_{KIDS}$, as compared with these other markers of ''endoscopic competence,'' is its ability to identify particular areas of strength and weakness, thus providing trainees with specific targeted feedback.

The $GiECAT_{KIDS}$ was designed for use as both a formative and a summative assessment tool, to monitor the development of an endoscopist's technical, cognitive, and integrative competencies in a standardized manner, throughout training, as he or she progresses from novice through the continuum to competent endoscopist. Formative assessment is intended to provide feedback to the learner to modify his or her thinking or behavior to improve and shape learning (8). The present study provides reliability and validity evidence to support use of the $GiECAT_{KIDS}$ in a formative manner throughout training to help identify specific areas of weakness and aid in the provision of targeted feedback. Formative assessment instruments should ideally foster meaningful, timely, task-focused, and goal-oriented feedback (8,27), something enabled by the task-specific nature of the $GiECAT_{KIDS}$. In addition, procedure-specific direct observation tools, such as the $GiECAT_{KIDS}$, support competency-based curricular designs because they facilitate the documentation of progress toward identified outcomes and provide a method to gather information necessary to make a case for the competence of a trainee (27). Although other markers of ''endoscopic competence'' such as terminal ileal intubation rate may reflect ability, they do not provide trainees with targeted feedback. Use of the $GiECAT_{KIDS}$ during training would allow endoscopic trainers to better identify specific skills that require attention such as loop reduction, knowledge of the informed consent process, or patient communication skills. Additional studies using aggregate outcome data from large numbers of pediatric endoscopists across North America are required to establish average normal learning curves of endoscopists' $GiECAT_{KIDS}$ scores to define the milestones along the learning curve and generate specific proficiency benchmarks for use in high-stakes summative assessment, certification, and recertification. Establishment of normal learning curves based on aggregated data would allow program directors to plot their fellows on such curves to quickly identify specific skills that are lacking and pinpoint trainees requiring remedial or specialized attention.

The present study had a number of limitations. First, because of the clinical nature of the assessments, it was not possible to blind the raters to the training status of the endoscopists. The $GiECAT_{KIDS}$ GRS anchors were deliberately aligned with the construct of developing clinical independence, a strategy that has shown to improve rater agreement and discrimination among trainees of varying abilities (28). Although the rating scales have well-defined criteria,

TABLE 6. Concurrent validity of total $GiECAT_{KIDS}$, global rating scale, and checklist components

Component of $GiECAT_{KIDS}$ scale	Correlation coefficient (Spearman ρ)			
	No. previous colonoscopies	Cecal intubation rate	Terminal ileal intubation rate	Physician global assessment of skill
Total $GiECAT_{KIDS}$ score*	0.91	0.82	0.82	0.95
Global rating scale score*	0.92	0.85	0.82	0.94
Checklist score*	0.84	0.77	0.80	0.89

$GiECAT_{KIDS}$ = Gastrointestinal Endoscopy Competency Assessment Tool for Pediatric Colonoscopy.
* Correlations significant ($P < 0.001$).

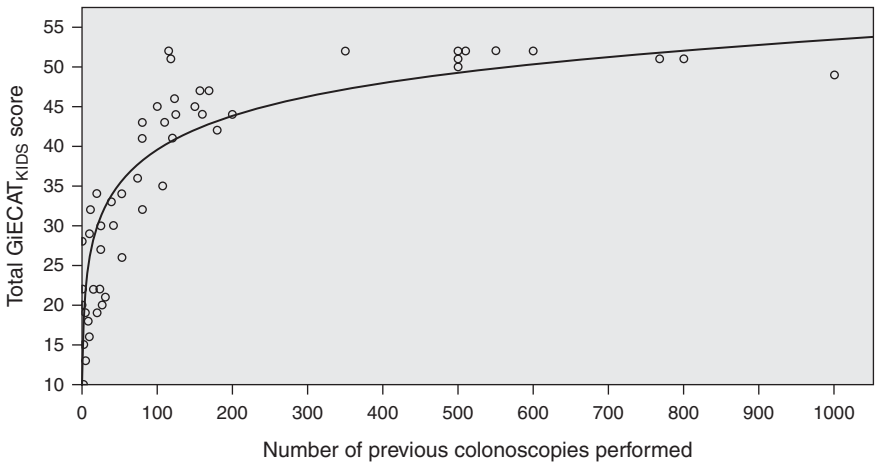

FIGURE 1. Scatterplot of total GiECAT$_{KIDS}$ scores versus number of previous procedures with line of best fit. GiECAT$_{KIDS}$ = Gastrointestinal Endoscopy Competency Assessment Tool for Pediatric Colonoscopy.

the raters may have been biased with the knowledge of the level of training. Future work is planned to assess reliability and validity of the GiECAT$_{KIDS}$ in the context of blinded expert video ratings because this would allow for determination of intrarater reliability and reduce potential bias. Second, clinical performance assessment has been criticized as endoscopists' awareness of being observed may create anxiety and lead to poor performance, resulting in appraisals that may not accurately reflect true performance ability (29). Finally, validity and reliability evidence is context specific, and thus further validation is required to determine suitability of the GiECAT$_{KIDS}$ for use within the simulated setting. Future work is also planned to compare the psychometric properties of the GiECAT$_{KIDS}$ with colonoscopy assessment tools developed within the adult context, such as the Mayo Colonoscopy Skill Assessment Tool (30).

The GiECAT$_{KIDS}$ assessment tool has several strengths. It is the first procedure-specific colonoscopy assessment tool that has been developed and validated within the pediatric context (12). Using Delphi methodology, scale items were selected by a panel of North American pediatric endoscopy experts, thus providing content-related validity evidence of the resultant tool and increasing transferability across institutions (15,31). The present study provides support for the use of the GiECAT$_{KIDS}$ by demonstrating evidence of the interrater and test–retest reliability and concurrent and discriminative validity of the tool within the clinical setting. In addition, the tool's strong psychometric properties afford evidence for the use of Delphi methodology for the creation of procedural evaluation tools reflective of international practice.

In summary, the present study provides evidence supporting the reliability and validity of the GiECAT$_{KIDS}$ for use by a single physician rater to assess performance of live pediatric colonoscopy within the clinical setting in the context of formative assessment to provide endoscopists with feedback during the course of their training. It has a user-friendly, logical structure and is easily administered within the clinical environment. In addition, it provides trainees with feedback about specific aspects of their performance. Integration of the GiECAT$_{KIDS}$ assessment tool into training has the ability to facilitate the provision of constructive formative feedback, aid in the identification of skills requiring remediation, provide a means to document trainees' progress over time, and help identify learners in difficulty. Before use in very-high-stakes assessment, such as credentialing, further research is required to ensure that a higher interrater reliability is achieved with raters who are formally trained. In addition, it is hoped that aggregate data derived from longitudinal use within pediatric gastroenterology training programs across North America can be used to generate average normal learning curves for pediatric endoscopists to help define milestones along the learning curve and aid in the establishment of minimal performance-based standards for competence in pediatric colonoscopy.

Acknowledgments: The authors thank Drs Brian Hodges and Dorcas Beaton for their insightful comments.

REFERENCES

1. Takahashi SG, Waddell A, Kennedy M, et al. Innovations, integration and implementation issues in competency-based education in postgraduate medical education. *https://www.afmc.ca/pdf/fmec/19_Glover Takahashi_Competency-based Education.pdf*. Published 2011. Accessed June 1, 2014.
2. Long DM. Competency-based residency training: the next advance in graduate medical education. *Acad Med* 2000;75:1178–83.
3. Leichtner AM, Gillis LA, Gupta S, et al. NASPGHAN guidelines for training in pediatric gastroenterology. *J Pediatr Gastroenterol Nutr* 2013;56 (suppl 1):S1–8.
4. Walsh CM, Ling SC, Khanna N, et al. Gastrointestinal Endoscopy Competency Assessment Tool: development of a procedure-specific assessment tool for colonoscopy. *Gastrointest Endosc* 2014;79:798–807.e5.
5. Royal College of Physicians and Surgeons of Canada. Objectives of training in the subspecialty of gastroenterology. *www.royalcollege.ca*. Published 2011. Accessed June 1, 2014.
6. Accreditation Council for Graduate Medical Education. ACGME program requirements for graduate medical education in pediatric gastroenterology. *http://www.acgme.org/acgmeweb/Portals/0/PFAssets/2013-PR-FAQ-PIF/332_gastroenterology_peds_07012013.pdf*. Published 2013. Accessed June 1, 2014.
7. Sidhu R, Grober E, Musselman L, et al. Assessing competency in surgery: where to begin? *Surgery* 2004;135:6–20.
8. Shute VJ. Focus on formative feedback. *Rev Educ Res* 2008;78:153–89.
9. Klasko SK, Cummings RV, Glazerman LR. Education resident data collection: do the numbers add up? *Am J Obstet Gynecol* 1995;172 (4 pt 1):1312–6.
10. Sedlack RE. Training to competency in colonoscopy: assessing and defining competency standards. *Gastrointest Endosc* 2011;74:355–66e1–2.
11. Van Der Vleuten CP. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract* 1996;1:41–67.

12. Walsh CM, Ling SC, Walters TD, et al. Development of the Gastrointestinal Endoscopy Competency Assessment Tool for Pediatric Colonoscopy ($GiECAT_{KIDS}$). *J Pediatr Gastroenterol Nutr* 2014;59:480–6.
13. Walsh CM. Assessment of competence in pediatric gastrointestinal endoscopy. *Curr Gastroenterol Rep* 2014;16:401.
14. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997; 84:273–8.
15. Cook DA, Zendejas B, Hamstra SJ, et al. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Health Sci Educ Theory Pract* 2014;19:233–50.
16. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420–8.
17. Streiner DL, Norman GR. Reliability. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 4th ed. New York: Oxford University Press; 2008.
18. Weir JP. Quantifying test–retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res* 2005;19:231–40.
19. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006; 119:166.e7–16.
20. Watkins M, Portney L. *Foundations of Clinical Research: Applications to Practice*. Upper Saddle River, NJ: Pearson Education; 2009.
21. Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ* 2004;38:1006–12.
22. Wright JG, Feinstein AR. Improving the reliability of orthopaedic measurements. *J Bone Joint Surg Br* 1992;74:287–91.
23. Gallagher AG, Ritter EM, Satava RM. Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. *Surg Endosc* 2003;17:1525–9.
24. Lurie SJ, Mooney CJ, Lyness JM. Measurement of the general competencies of the accreditation council for graduate medical education: a systematic review. *Acad Med* 2009;84:301–9.
25. Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: rethinking the etiology of rater errors. *Acad Med* 2011;86 (10 suppl):S1–7.
26. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association; 1999.
27. Donato AA. Direct observation of residents: a model for an assessment system. *Am J Med* 2014;127:455–60.
28. Crossley J, Johnson G, Booth J, et al. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. *Med Educ* 2011;45:560–9.
29. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med* 2003;15:270–92.
30. Sedlack RE. The Mayo Colonoscopy Skills Assessment Tool: validation of a unique instrument to assess colonoscopy skills in trainees. *Gastrointest Endosc* 2010;72:1125–331133.e1–3.
31. De Villiers MR, de Villiers PJT, Kent AP. The Delphi technique in health sciences education research. *Med Teach* 2005;27:639–43.