Is there a Valid and Reliable Assessment of Diagnostic Knee Arthroscopy Skill?

Tyson Olson BS, Ryan Koehler BS, Aaron Butler BS,
Simon Amsdell MD, Gregg Nicandri MD

Received: 23 July 2012 / Accepted: 4 December 2012 / Published online: 20 December 2012
© The Association of Bone and Joint Surgeons® 2012

Abstract

Background The Basic Arthroscopic Knee Skill Scoring System (BAKSSS) has construct validity as an objective measure of arthroscopic proficiency when used to assess the task of performing arthroscopic meniscectomies on cadaver knees. The reliability of this instrument is unknown.

Questions/purpose We asked whether (1) a simple modification of the BAKSSS would show construct validity similar to that in the initial BAKSSS study, (2) this assessment would be reliable, and (3) this assessment could be used as a high-stakes pass or fail test.

Methods Twenty-three orthopaedic residents performed diagnostic knee arthroscopies on cadaveric knee specimens. Their competency was assessed by three live raters using the modified BAKSSS. Interrater reliability was assessed by comparing the scores given by each rater to each subject.

Results The modified BAKSSS showed construct validity with junior residents achieving lower scores (mean score, 20) than senior residents (mean score, 33). The modified BAKSSS had an interrater reliability of kappa = 0.685–0.852. The modified BAKSSS had a kappa = 0.543 when used as a proficiency test for diagnostic arthroscopy.

Conclusions The modified BAKSSS is useful for assessing diagnostic knee arthroscopy proficiency. Future scoring systems should be designed to be generalizable so they can be applied to multiple procedures without the need for modification, allow for video-based assessment, and must be rigorously tested for reliability and other types of validity (eg, face validity, content validity, and criterion-related validity).

Clinical Relevance Having a valid and reliable assessment of basic arthroscopic procedures may allow educators to more adequately evaluate individual residents and the effectiveness of various training modalities.

Each author certifies that he or she has no commercial associations (eg, consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article.

All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research* editors and board members are on file with the publication and can be viewed on request.

Clinical Orthopaedics and Related Research neither advocates nor endorses the use of any treatment, drug, or device. Readers are encouraged to always seek additional information, including FDA-approval status, of any drug or device prior to clinical use. Each author certifies that his or her institution approved the human protocol for this investigation, that all investigations were conducted in conformity with ethical principles of research, and that informed consent for participation in the study was obtained.

T. Olson, R. Koehler, A. Butler, S. Amsdell, G. Nicandri (✉)
Department of Orthopaedics and Rehabilitation, University
of Rochester Medical Center, 601 Elmwood Avenue,
Box 665, Rochester, NY, USA
e-mail: Gregg_Nicandri@urmc.rochester.edu

Introduction

Traditionally, the surgical competency of residents has been assessed by an attending educator's observations in the operating room, subjective end-of-rotation evaluations, and case logs. There has been a demand for more objective measures of surgical technical skill, and these measures have been used successfully by multiple disciplines [5, 10, 11, 13, 17, 19–21].

Similar tools have been studied for use in orthopaedic assessment [3, 6–8, 14–16]. Insel et al. [8] developed the Basic Arthroscopic Knee Skill Scoring System (BAKSSS) as an objective measure to evaluate the performance of arthroscopic meniscectomy on a cadaver knee in the bioskills laboratory. Establishing the reliability and validity of such

assessment tools is critical to ensure accurate trainee evaluation. Validation of an assessment tool can be established by proving content validity, construct validity, predictive validity, test-retest reliability, internal consistency, and interrater reliability [18]. Insel et al. [8] established the content validity and construct validity of their instrument; however, they did not assess the reliability of the BAKSSS. In our opinion, a scoring system that could be used to evaluate diagnostic knee arthroscopy, one of the first skills trainees learn, would be valuable to orthopaedic educators. In search of such a tool, we reviewed the BAKSSS and thought it could be easily modified to accomplish this goal.

We asked whether (1) a simple modification of the BAKSSS would show construct validity similar to that in the initial BAKSSS study, (2) this assessment would be reliable, and (3) this assessment could be used as a high-stakes pass or fail test.

Materials and Methods

We sent a recruitment email and information letter to all 34 orthopaedic residents in a single orthopaedic program in

November 2011. Twenty-three of the 34 residents (13 juniors and 10 seniors) self-enrolled in this study and completed a short online demographics survey. Demographic questions pertained to sex, handedness, postgraduate year (PGY), and the number of cases that were logged in the Accreditation Council for Graduate Medical Education weblog under “arthroscopy of the knee/femur” (Table 1). Participants who completed the study and demographics survey were given \$10 gift cards as compensation. This study was approved by the research subjects review board.

A power analysis was conducted for a β of 80% and an α of 0.05. Based on the original data from Insel et al. [8], a clinically important difference of 10 points on the modified BAKSSS and a SD of 7 for each group were assigned. The power analysis suggested a minimum of eight subjects per group (junior versus senior) would be required to identify meaningful differences.

The simulated procedures were performed at the University of Rochester's Kenneth Dehaven Arthroscopic Skills Laboratory. Two arthroscopic stations, one for right cadaveric knees and the other for left cadaveric knees, were set up in the skills laboratory. Residents were assigned to complete diagnostic arthroscopies of either the right or left

Table 1. Demographic data and modified BAKSSS score for each study participant

Participant	Sex	PGY	Hand	Technical skills training hours	Number of knee cases	Rater 1	Rater 2	Rater 3	Mean BAKSSS
1	M	1	R	5	20	22	18	27	22.33
2	F	1	R	0	50	19	14	21	18.00
3	M	1	R	16	0	24	28	21	24.33
4	M	1	R	0	0	15	14	21	16.67
5	M	1	R	6	0	24	19	27	23.33
6	M	1	R	55	52	14	16	31	30.33
7	M	2	R	7	0	11	16	20	11.75
8	M	2	R	0	50	26	29	29	28.00
9	F	2	R	6	12	13	11	15	13.00
10	F	2	R	52	12	23	22	28	24.33
11	M	2	R	1	37	18	19	14	17.00
12	M	2	R	6	15	29	25	29	27.67
13	M	2	R	15	6	15	15	22	17.33
14	M	3	R	20	140	30	28	36	31.33
15	F	4	R	65	171	36	29	36	33.67
16	M	4	R	142	89	37	32	38	35.67
17	M	4	R	66	120	34	36	33	34.33
18	M	4	R	49	80	33	28	31	30.67
19	F	4	R	22	80	33	30	28	30.33
20	M	5	R	20	120	38	37	37	37.33
21	M	5	R	19	192	35	36	31	34.00
22	F	5	R	26	225	29	31	34	31.33
23	M	5	R	9	172	37	39	30	35.33

PGY = Postgraduate year; BAKSSS = Basic Arthroscopic Knee Skill Scoring System.

knee, based on the date of the session they completed the study. Knee assignments were alternated from session to session and resulted in nine residents assigned to perform diagnostic knee arthroscopies on the right knee and 14 residents assigned to the left knee.

Before entering the laboratory, participants were asked to view a 4-minute video example of a trained surgeon performing diagnostic arthroscopy of the knee. In the laboratory, an attending surgeon (GN) was present at the arthroscopic station to assist the residents if they were unable to complete the task and to provide feedback after completion. Each participant's performance was assessed using the modified BAKSSS by three raters (one attending surgeon [GN] and two medical students [RK, TO]).

The BAKSSS consists of a 10-item, task-specific checklist and a 10-item global rating scale [8]. The global rating scale is a Likert-type scale, with a possible range of scores from 1 (lowest level of performance) to 5 (highest level of performance). Insel et al. [8] created weighted descriptors for scores 1, 3, and 5 on all 10 items. The original BAKSSS also used a task-specific checklist to assess 10 tasks determined to be important for performing diagnostic arthroscopy and partial meniscectomy. For this study, the task-specific checklist was not used, as global rating scales alone reportedly have better psychometric properties than checklists or a combination of a checklist and global rating scale [1]. One item, dissection, was omitted from the global rating scale. To limit the cost of the study, the same two cadaveric specimens were used for the entire study. As a result, all arthroscopic portals had to be made previously, and no other dissection was performed; therefore, this category could not be rated in this setting. Participants were assessed using this nine-item, modified BAKSSS global rating scale, including the parameters of instrument handling, depth perception, bimanual dexterity, flow of operation, knowledge of instruments, efficiency, knowledge of procedure, autonomy, and quality of the final product (Appendix 1).

The raters consisted of one attending surgeon and two medical students, each with previous experience using the BAKSSS global rating scale. All three raters simultaneously rated the performance of each participant in the skills laboratory, with each rater blinded to the scores given by the other raters. One medical student was blinded to the year in training of the residents. In keeping with the original BAKSSS study design, a second unblinded medical student also was used as an evaluator. The attending surgeon was not blinded. All three raters underwent rater training, which consisted of one session in which they were instructed on the intraarticular anatomy using a surgical video, a demonstration of diagnostic arthroscopy on one of the cadaveric specimens by a fellowship-trained sports medicine subspecialty physician, and instructions regarding how to use the BAKSSS to rate these videos. This

training session was similar to that described in other published studies that evaluated analogous assessments [8, 19]. To evaluate the modified BAKSSS as a proficiency assessment of diagnostic knee arthroscopic surgical skill, a minimal proficiency threshold was arbitrarily set as a total modified BAKSSS score of 30 of 45. This would be equivalent to one better than the average score (3) in each category, plus a score of 5 for autonomy. Residents were given a pass for a score of 30 or greater and a fail for a score less than 30.

A weighted Kappa statistic (κ) was used to assess interrater reliability between each of the three raters [4, 9]. A weighted Kappa statistic also was used to assess agreement between Rater 2 and Rater 3 for each domain of the modified BAKSSS. Residents were grouped into junior (PGYs-1 and 2) and senior (PGYs-3-5) resident categories, and a Student's *t*-test was used to compare the mean modified BAKSSS score for each group. A CI was set at 95%. The total pass rate by all raters for each PGY was determined. Kappa statistics were used to assess agreement between the raters when BAKSSS was used as a pass or fail test. Statistical analysis was performed using MedCalc software v12.1.4 (MedCalc, Mariakerke, Belgium).

Results

The modified BAKSSS global rating scale showed construct validity for assessment of diagnostic knee arthroscopy. Junior resident scores (mean score, 20) were significantly lower than senior resident scores (mean score, 33) ($p < 0.001$) (Fig. 1).

The interrater reliability of the modified BAKSSS global rating scale total score varied between $K = 0.69$ to

Fig. 1 The mean modified BAKSSS global rating scale scores of junior (PGYs 1-2) residents were lower ($p < 0.001$) than those of the senior (PGYs 3-5) residents: 20 and 33, respectively.

Table 2. Level of agreement for each domain between Raters 2 and 3

Domain	Kappa
Autonomy	0.41
Bimanual dexterity	0.44
Procedure knowledge	0.47
Instrument handling	0.48
Quality	0.54
Flow	0.57
Depth perception	0.60
Efficiency	0.77
Instrument knowledge	0.81

$K = 0.85$. For each individual domain, the κ values for the two medical student raters (Raters 2 and 3) ranged from 0.41 to 0.81 (Table 2).

When the modified BAKSSS was used as a pass or fail test, the interrater reliability between Raters 2 and 3 was only 0.54.

Discussion

The competency-based assessment of surgical technical skill is becoming increasingly important [2, 12]. After reviewing the BAKSSS, we thought that this tool could be easily modified to measure a surgeon's arthroscopic proficiency for one of the first skills that new trainees learn; that of diagnostic knee arthroscopy. In this investigation, we asked whether (1) this simple modification of the BAKSSS would show construct validity similar to that in the original BAKSSS study, (2) this assessment would be reliable, and (3) this assessment could be used as a high-stakes pass or fail test.

Our study has several limitations. First, similar to the initial BAKSSS validation study [8], we used live raters which makes blinding difficult. Knowing that this was a substantial limitation to the original BAKSSS study, we did ensure that one of the raters was strictly blinded. The level of agreement for the two medical student raters (Rater 2 unblinded and Rater 3 blinded) was $\kappa = 0.685$ leading us to believe that the assessment was sufficiently reliable for our purposes. Second, the threshold for passing (minimal proficiency) was arbitrarily chosen at 30. This number seemed to have the most face validity as it indicates at least an average score for each of the nine domains assessed and a perfect score for autonomy which we thought was an important indicator of minimal proficiency. The pass rate as determined by Rater 3 when the threshold for passing was set at 30 was 43%. Pass rates were no different when similar but equally arbitrary scores of 29 (52%) or 31

(43%) were used ($p = 0.75$ and $p = 0.77$ respectively). Third, the raters used in this study had differing levels of arthroscopic expertise (two medical students and one attending orthopaedist). This seemed to have a minimal effect, as there was no difference between average BAKSSS scores given by the attending surgeon and the medical students. This may not be the case, as the complexity of the procedures being evaluated increase to a point where it may be necessary that raters have more experience performing the procedures themselves. This study was designed to closely mimic the original BAKSSS, and medical student raters, experienced in using the BAKSSS, (in addition to the attending physician) were used. Fourth, we were limited because we were modifying an existing scale and perceived flaws in that scale were carried forward to this study. For example, it can be questioned whether domains that primarily assess content knowledge should be included on an assessment of technical skill. Although it is necessary for residents to know their surgical instruments and procedures, these content knowledge domains can be assessed in other ways. Although this limitation does not influence the overall outcomes of our study it is interesting that deleting the cognitive domains "knowledge of instruments" and "knowledge of specific procedure" would have decreased the overall reliability of the BAKSSS. Fifth, owing to the financial constraints of the study, only two cadaveric specimens were provided and all residents used the same portals. We therefore could not evaluate portal placement which is an integral part of any arthroscopic procedure. Despite this shortcoming however, the modified BAKSSS was able to discriminate proficiency between junior and senior residents. Finally, we tested only construct validity. We did not assess face validity, content validity, or criterion-related validity. These would need to be the subject of further study.

Similar to Insel et al. [8], we were able to show the construct validity of the modified BAKSSS, as residents with less experience scored much lower than those with more experience. Our residents, however, did not show the same gradual improvement in their BAKSSS scores with each postgraduate year in training, as was seen in the initial validation study of the BAKSSS [8] (Fig. 2). Since diagnostic knee arthroscopy is a relatively simple procedure, it likely has a shorter learning curve than that of arthroscopic partial meniscectomy. For simple tasks, global rating scales, such as the BAKSSS, have a harder time identifying differences between groups of trainees because of the "ceiling effect" [1]. In our program, most residents are not exposed to an extensive amount of arthroscopy until the third postgraduate year, when the experience is quite intensive with a high volume of cases. This may explain the similarity between group scores for junior (PGY-1 and PGY-2) and senior residents (PGY-3, PGY-4, and PGY-5).

The interrater reliability of the BAKSSS has not been assessed before. Without adequate interrater agreement these types of tools are not as useful. Poor interrater agreement increases the time commitment required of the rater, as one individual is needed to rate all subjects. In addition, if there are considerable variances in rater scores, it is not possible to accurately compare results from multiple studies where different raters are used. Worse, if more than one rater is used in a study, it could result in incorrectly accepting or rejecting the study hypothesis if rater scores show considerable variability. The interrater reliability for our modified BAKSSS global rating scale varied between 0.685 to 0.857 which is consistent with similar studies (Table 3). The agreement however was lower for many of the individual domains (Table 2). Future scoring systems should focus on using descriptors designed to eliminate variability in rater scores for each domain to maximize the reliability of the instrument.

It is possible that a scoring system similar to BAKSSS may be used as a high-stakes test. For this purpose, the construct seems valid, as pass rates increased as the PGY of

the tested resident also increased (Fig. 3). The interrater reliability however was only 0.543. Of the 23 residents, only five (two were PGY-4, and three were PGY-5) received passing scores (30 or greater) from all raters. Ten (all PGYs-4 and 5 residents, and one PGY-3 resident) received passing scores from at least one rater. The pass rate, therefore, varied from 21% (five of 23) to 43% (10 of 23), depending on which rater was performing the examination. We believe this variability was too great at this point for the modified BAKSSS to be considered feasible for high-stakes testing.

We believe an ideal assessment tool for high-stakes testing would be generalizable to multiple arthroscopic procedures without the need for modification, (this would eliminate the need to rigorously validate multiple tools specific to each procedure), would allow for video-based assessments (this would facilitate the blinding of assessors), and would show a high level of interrater reliability.

We believe that the modified BAKSSS shows construct validity for assessment of diagnostic knee arthroscopy

Fig. 2 The mean modified BAKSSS global rating scale scores by resident PGY are shown. After PGY2, the mean scores increased as PGY of the resident increased.

Fig. 3 Pass rates of subjects by residency PGY when using the modified BAKSSS global rating scale as a pass or fail test are shown. After PGY 2, the pass rates increased as PGY of the resident increased.

Table 3. Comparison of the modified BAKSSS to other global rating scales

Study	Assessment tool	Subjects	Raters	Construct validity	Reliability
Martin et al. [10]	OSATS global rating scale (basic surgical skill)	20	48	Yes	0.70 to 0.72 [#]
VanHeest et al. [16]	OSATS global rating scale (carpal tunnel release)	28	2	Yes	0.69 [#]
Vassiliou et al. [19]	GOALS direct observation (cholecystectomy)	21	2	Yes	0.89 [#]
Vassiliou et al. [20]	GOALS video observation (cholecystectomy)	10	4	Yes	0.68 [#]
Insel et al. [8]	BAKSSS direct observation (partial meniscectomy)	68	1	Yes	Not available
Current study	Modified BAKSSS direct observation (diagnostic knee arthroscopy)	23	3	Yes	0.69-0.85 ^{##}

[#] ICC = intraclass correlation coefficient; ^{##}Kappa; BAKSSS = Basic Arthroscopic Knee Skill Scoring System; OSATS = Objective Structured Assessment of Technical Skills; GOALS = Global Operative Assessment of Laparoscopic Skill.

skill. When compared with other surgical assessments, it has similar levels of interrater agreement; however, improvement in this area would make the instrument more useful. In its current state, it is not feasible to use as a high-stakes test. The development of future assessments which are generalizable to multiple procedures without the need for modification, allow for video-based assessments, and are rigorously tested for reliability may better facilitate high-stakes testing.

Acknowledgments We thank Rachel Johnston for assistance in formatting this manuscript.

Appendix 1: Modified BAKSSS Global Rating Scale

Instrument handling

- 1- Repeatedly makes tentative or awkward movements with instruments
- 2-
- 3- Competent use of instruments, although occasionally appeared stiff or awkward
- 4-
- 5- Fluid moves with instruments and no awkwardness

Depth perception

- 1- Constantly overshoots target, slow to correct
- 2-
- 3- Some overshooting or missing of target
- 4-
- 5- Accurately directs instruments in the correct plane to target

Bimanual dexterity

- 1- Noticeably awkward with nondominant hand, poor coordination between hands
- 2-
- 3- Uses both hands but does not maximize interaction between hands
- 4-
- 5- Expertly uses both hands in complementary manner to provide optimum performance

Flow of operation and forward planning

- 1- Frequently stopped operating or needed to discuss next move
- 2-
- 3- Demonstrated ability for forward planning with steady progression of operative procedure
- 4-
- 5- Obviously planned course of operation with effortless flow from one move to the next

Knowledge of instruments

- 1- Frequently asked for the wrong instrument or used inappropriate instrument
- 2-
- 3- Knew the names of most instruments and used appropriate instrument for the task
- 4-
- 5- Obviously familiar with the instruments required and their names

Efficiency

- 1- Many unnecessary, inefficient movements, constantly changing focus or persisting without progress
- 2-
- 3- Slow but planned movements are reasonably organized with few unnecessary or repetitive movements
- 4-
- 5- Confident, clear economy of movement and maximum efficiency.

Knowledge of specific procedure

- 1- Deficient knowledge, needed specific instruction at most operative steps
- 2-
- 3- Knew all important aspects of the operation
- 4-
- 5- Demonstrated familiarity with all aspects of the operation

Autonomy

- 1- Unable to complete entire task, even with verbal guidance
- 2-
- 3- Able to complete task safely with moderation guidance
- 4-
- 5- Able to complete task independently without prompting

Quality

- 1- Very poor
- 2-
- 3- Competent
- 4-
- 5- Clearly superior

References

1. Alvand A, Auplish S, Gill H, Rees J. Innate arthroscopic skills in medical students and variation in learning curves. *J Bone Joint Surg Am.* 2011;93:e115(111–119).

2. Atesok K, Mabrey JD, Jazrawi LM, Egol KA. Surgical simulation in orthopaedic skills training. *J Am Acad Orthop Surg.* 2012;20:410–422.
3. Ceponis PJ, Chan D, Boorman RS, Hutchison C, Mohtadi NG. A randomized pilot validation of educational measures in teaching shoulder arthroscopy to surgical residents. *Can J Surg.* 2007;50:387–393.
4. Doyle JD, Webber EM, Sidhu RS. A universal global rating scale for the evaluation of technical skills in the operating room. *Am J Surg.* 2007;193:551–555; discussion 555.
5. Gumbs AA, Hogle NJ, Fowler DL. Evaluation of resident laparoscopic performance using global operative assessment of laparoscopic skills. *J Am Coll Surg.* 2007;204:308–313.
6. Howells NR, Gill HS, Carr AJ, Price AJ, Rees JL. Transferring simulated arthroscopic skills to the operating theatre: a randomised blinded study. *J Bone Joint Surg Br.* 2008;90:494–499.
7. Hoyle AC, Whelton C, Umaar R, Funk L. Validation of a global rating scale for shoulder arthroscopy: a pilot study. *Shoulder Elbow.* 2012;4:16–21.
8. Insel A, Carofino B, Leger R, Arciero R, Mazzocca AD. The development of an objective model to assess arthroscopic performance. *J Bone Joint Surg Am.* 2009;91:2287–2295.
9. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159–174.
10. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, Brown M. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg.* 1997;84:273–278.
11. Mitchell EL, Arora S, Moneta GL. Ensuring vascular surgical training is on the right track. *J Vasc Surg.* 2011;53:517–525.
12. Pedowitz RA, Marsh JL. Motor skills training in orthopaedic surgery: a paradigm shift toward a simulation-based educational curriculum. *J Am Acad Orthop Surg.* 2012;20:407–409.
13. Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative “bench station” examination. *Am J Surg.* 1997;173:226–230.
14. Sonnadara RR, Van Vliet A, Safir O, Alman B, Ferguson P, Kraemer W, Reznick R. Orthopedic boot camp: examining the effectiveness of an intensive surgical skills course. *Surgery.* 2011;149:745–749.
15. Tashiro Y, Miura H, Nakanishi Y, Okazaki K, Iwamoto Y. Evaluation of skills in arthroscopic training based on trajectory and force data. *Clin Orthop Relat Res.* 2009;467:546–552.
16. Van Heest A, Putnam M, Agel J, Shanedling J, McPherson S, Schmitz C. Assessment of technical skills of orthopaedic surgery residents performing open carpal tunnel release surgery. *J Bone Joint Surg Am.* 2009;91:2811–2817.
17. VanHeest A, Kuzel B, Agel J, Putnam M, Kalliaainen L, Fletcher J. Objective structured assessment of technical skill in upper extremity surgery. *J Hand Surg Am.* 2012;37:332–337 337.e4.
18. Van Nortwick SS, Lendvay TS, Jensen AR, Wright AS, Horvath KD, Kim S. Methodologies for establishing validity in surgical simulation studies. *Surgery.* 2010;147:622–630.
19. Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondre K, Stanbridge D, Fried GM. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg.* 2005;190:107–113.
20. Vassiliou MC, Feldman LS, Fraser SA, Charlebois P, Chaudhury P, Stanbridge DD, Fried GM. Evaluating intraoperative laparoscopic skill: direct observation versus blinded videotaped performances. *Surg Innov.* 2007;14:211–216.
21. Vassiliou MC, Kaneva PA, Poulouse BK, Dunkin BJ, Marks JM, Sadik R, Sroka G, Anvari M, Thaler K, Adrales GL, Hazey JW, Lightdale JR, Velanovich V, Swanstrom LL, Mellinger JD, Fried GM. Global Assessment of Gastrointestinal Endoscopic Skills (GAGES): a valid measurement tool for technical skills in flexible endoscopy. *Surg Endosc.* 2010;24:1834–1841.