

# Evaluation of surgical fellows' laparoscopic performance using Global Operative Assessment of Laparoscopic Skills (GOALS)

Nancy J. Hogle · Yi Liu · R. Todd Ogden ·  
Dennis L. Fowler

Received: 28 May 2013 / Accepted: 6 November 2013 / Published online: 11 January 2014  
© Springer Science+Business Media New York 2014

## Abstract

**Background** The acquisition of technical skills is one of the fundamental goals of postgraduate surgical training; however, validation and utilization of objective tools to assess the technical skills of trainees remains elusive. The objectives of this project are to develop models to identify predictive factors for fellow performance, validate the Global Operative Assessment of Laparoscopic Skills (GOALS) as an assessment tool for laparoscopic skills, and to define the learning curve for complex laparoscopic gastrointestinal surgery.

**Methods** Using previously recorded data from a centralized database of the Fellowship Council, we analyzed the voluntarily submitted performance scores of surgical fellows for three complex laparoscopic gastrointestinal operations: Roux-en-Y gastric bypass, LapBand placement, and Nissen fundoplication. We analyzed previous experience with complex cases, previous experience with the same type of case, case difficulty, and time of year in the fellowship as potential predictors of performance. Performance scores throughout the fellowship year were graphed to create learning curves for overall performance and each of five domains of performance.

**Results** Ninety-eight performance assessments were submitted for 31 unique fellows. Overall performance ( $p < 0.01$ ), bimanual dexterity ( $p < 0.01$ ), efficiency

( $p < 0.01$ ), and autonomy ( $p < 0.01$ ) all improved significantly throughout the course of the fellowship year. Performance in the domains of depth perception and tissue handling improved, but the improvement did not reach statistical significance. Three predictor variables were significantly related to performance scores.

**Conclusions** This study documents that GOALS is able to differentiate novice fellows from graduating fellows and established construct validity. Models developed and tested confirmed that previous experience, case difficulty, and length of time as a fellow impacted performance.

**Keywords** Construct validity · GOALS · Surgery fellow evaluation · Validation · Laparoscopic surgery · Laparoscopic skills assessment

The acquisition of technical skills is one of the fundamental goals of postgraduate surgical training. Nearly 80 % of graduating chief residents in general surgery apply for a fellowship, and nearly 30 % apply to the Fellowship Council for a fellowship (advanced gastrointestinal/minimally invasive surgery, bariatric surgery, hepatobiliary surgery, etc.). This is an indication that these residents feel the need for more training, specifically training to acquire expert technical skills. Although both faculty and residents perceive this need, an objective method to assess the technical skills of trainees remains elusive. Program directors have never had valid, objective, reliable tools with which to measure the technical performance of residents and fellows and, therefore, have never been able to assess objectively which trainees are proficient and which trainees need additional technical training.

Traditionally, surgeons developed their technical skills through a process of protracted exposure to a supervised,

---

N. J. Hogle (✉) · D. L. Fowler  
Department of Surgery, Columbia University College  
of Physicians and Surgeons, 650 W 168th Street,  
BB 210, New York, NY 10032, USA  
e-mail: nancyhogle@gmail.com

Y. Liu · R. T. Ogden  
Department of Biostatistics, Columbia University Mailman  
School of Public Health, New York, NY, USA

graded operative experience. The skills required to complete an operation are the result of applying a specific combination of abilities to the task of the operation, and the assessment of operative performance measures the overall efficiency of completing the task by measuring the component abilities required to complete the task [1]. Procedural training in current practice often is unsystematic and unstructured; it is typically based on the random need to perform various surgical procedures. Surgical educators have worked for decades to improve the methods by which surgical education is delivered to trainees [2, 3]. However, validated teaching methods have not been integrated into clinical education, in part because validated tools with which to assess trainee performance did not exist [3]. Surgical fellows and their preceptors spend a large amount of time demonstrating and observing technical skills in the operating room, but traditional methods of assessment of technical skills rely upon subjective evaluations by senior staff members, case records, and morbidity and mortality rates [4], none of which have been shown to be valid or reliable.

Surgical educators need validated assessment tools to assess surgical competency during training to ensure that graduates have acquired sufficient skill to be safe and effective in independent practice [5–14]. Although surgeons have developed a few tools to evaluate performance and some have been validated for specific procedures, surgical educators and certifying organizations have not endorsed or mandated a comprehensive, objective,

validated method to measure technical performance. The basic reason to assess performance is to guide training to the level of proficiency; however, public demand, reduced resident work hours [15], and regulatory mandates have all increased the sense of urgency to develop and implement validated, objective assessment tools that will reliably measure technical performance.

### Global Operative Assessment of Laparoscopic Skills (GOALS)

The GOALS evaluates laparoscopic surgery performance in five domains (depth perception, bimanual dexterity, efficiency, tissue handling, and autonomy). Each domain is scored with a rating from 1 to 5. A descriptive anchor is provided for scores of 1, 3, and 5 for each domain (Table 1). A total score for each operation is the sum of the scores from the five domains and may be considered an overall assessment of a trainee's performance. GOALS was first validated for dissection of the gallbladder from the liver bed [16]. Subsequently, GOALS has been validated as an assessment tool for laparoscopic cholecystectomy and laparoscopic appendectomy but has not previously been validated for assessing performance during more complex operations [17]. Chang and colleagues also demonstrated construct validity and inter-rater reliability using GOALS to evaluate video-recordings of laparoscopic cholecystectomy performed by a novice and by an expert [18].

**Table 1** Basic GOALS for Nissen fundoplication, gastric band, and gastric bypass

| Domains            | 1  | 2 | 3  | 4 | 5  |
|--------------------|--|---|--|---|--|
| Depth perception   | Constantly overshooting target, hits backstop, wide swings, slow to correct  |   | Some overshooting or missing plane but corrects quickly  |   | Accurately directs instruments in correct plane to target  |
| Bimanual dexterity | Use of 1 hand, ignoring non-dominant hand, poor coordination between hands   |   | Use of both hands but does not optimize interactions between hands to facilitate conduct of operation  |   | Expertly utilizes both hands in a complementary manner to provide optimal working exposure   |
| Efficiency         | Uncertain, much wasted effort, many tentative motions, constantly changing focus of operation, or persisting at a task without progress          |   | Slow, but planned and reasonably organized   |   | Confident, efficient and safe conduct of operation, maintaining focus on component of procedure until better done by another approach                              |
| Tissue handling    | Rough, tears tissue by excessive traction, injures adjacent structures, poor control of coagulation device (recoil), grasper frequently slip off |   | Handles tissues reasonably well, with some minor trauma to adjacent tissues e.g. coagulation of liver, causes unnecessary liver bleeding, occasional slipping of grasper |   | Handles tissues very well with appropriate traction on tissues and negligible injury of adjacent structures. Uses energy sources appropriately but not excessively |
| Autonomy           | Unable to complete entire procedure, even in a straightforward case and with extensive verbal guidance   |   | Able to complete operation safely with moderate prompting  |   | Able to complete operation independently without prompting   |

More recently, investigators have developed and validated tools designed to evaluate performance during groin hernia repair (GOALS-GH) [19] and incisional hernia repair (GOALS-IH) [20, 21]. The lack of consensus regarding a validated assessment tool for use in most complex laparoscopic surgery cases led to the purpose of this study. We hypothesize that GOALS will differentiate novice fellows from graduating fellows and, thereby, establish construct validity for GOALS when used to evaluate performance during complex laparoscopic gastrointestinal surgery. Additionally, we hypothesize that there is a correlation between the GOALS scores for fellows and (1) total previous experience with complex laparoscopic operations, (2) previous experience with the procedure under evaluation, (3) difficulty of operation, and (4) length of time as a fellow.

## Materials and methods

The Fellowship Council provided a data set that included all voluntarily reported performance scores for fellows between June 2010 and November 2011. In addition to the performance scores, the data set included the percent of the case performed by the fellow, the name and date of the procedure, previous experience with complex cases, previous experience with the case being reported, and the difficulty of the case. With permission from the Fellowship Council and approval by the Institutional Review Board, the data were stratified by case type and then analyzed to identify predictors of improved performance.

### Statistical analysis

Some basic descriptive statistics were first calculated, including mean and standard deviation of the scores for each domain for each quarter during the fellowship year. In order to determine which of the various available factors (number of previous complex procedures performed, number of the same type of procedure performed, case difficulty, and total time in the fellowship program) were related to scores, we applied a linear mixed effects model

to the data. Each of the available factors is treated as a fixed effect and the fellow is included as a random effect. This allows us to model fellow-to-fellow variability and to account for correlation among different scores for a given subject. To allow for apparent nonlinearity of the learning curve, we log-transformed the time in the program. For all analyses, R version 2.15.2 ([www.rproject.org](http://www.rproject.org)) for Windows was used.

## Results

The performance of each of 31 unique fellows during 98 complex laparoscopic operations was assessed using the GOALS tool for three types of operations: laparoscopic Nissen fundoplication ( $n = 27$ ), laparoscopic adjustable gastric band ( $n = 17$ ), and laparoscopic Roux-en-Y gastric bypass ( $n = 54$ ; Table 2). Of the 31 fellows, 14 (45.2 %) had only one set of performance scores available. The mean scores for each quarter for each domain increased throughout the fellowship year (Fig. 1).

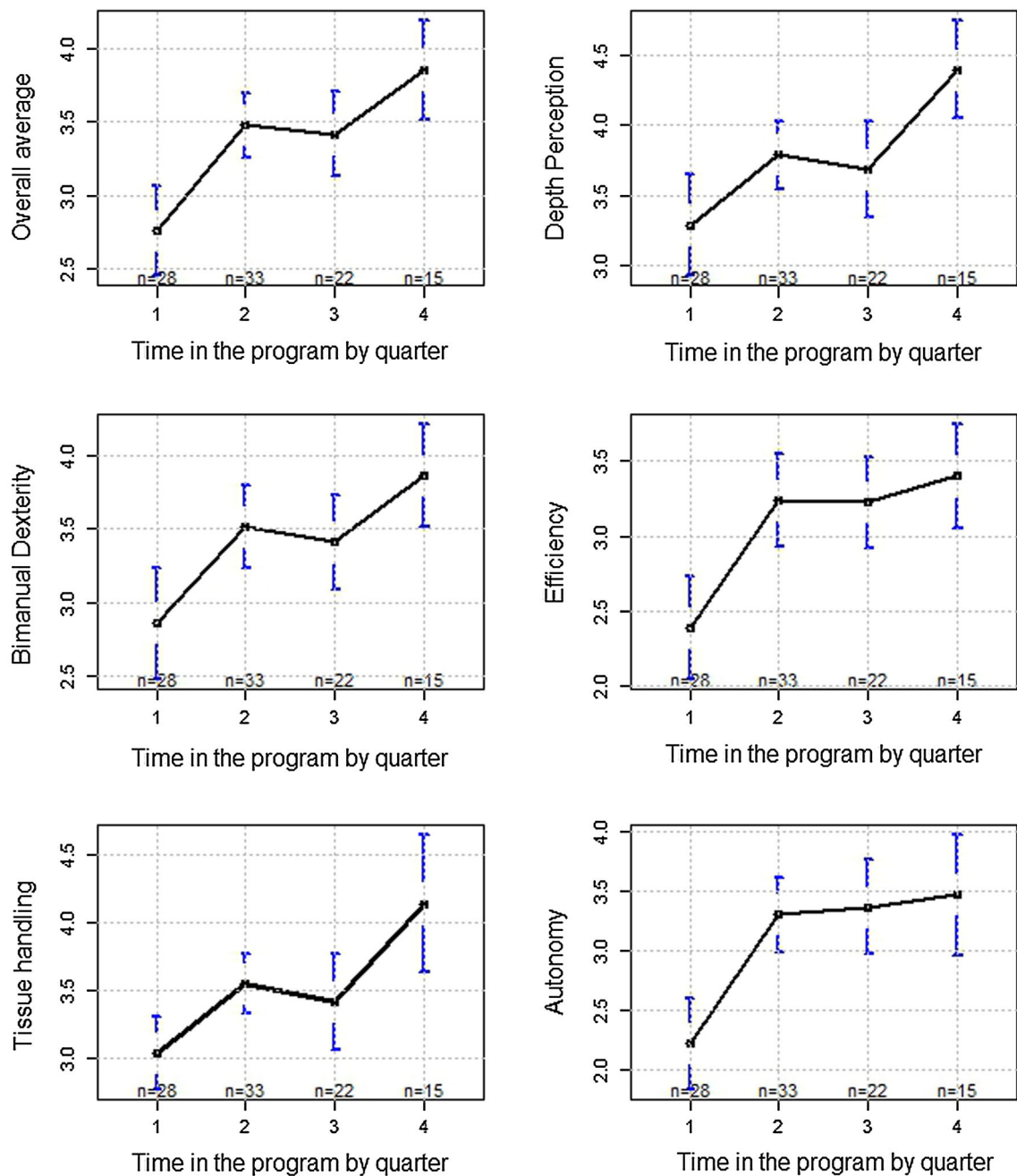
Table 3 displays the results of analyses using three different models. The mean of performance scores is the outcome variable in each model. On the left of the table (“Separate Models”) are the results from fitting a separate linear mixed effects model for each fixed effect predictor. In the middle of Table 3 (“Full model”) are the results of the analysis in which all four factors are included as fixed effects in a single model. On the right of Table 3 (“Final model”) are the results from the model giving the lowest Bayesian Information Criterion [22]. In this model, each of the three predictor variables is significantly related to the outcome.

Figure 2 shows all of the raw data and a calculated, estimated average learning curve for overall performance and for performance in each of the domains using the Final Model. The learning curves for overall performance, bimanual dexterity, efficiency, and autonomy demonstrated statistically significant learning during the fellowship year. Although the learning curves for depth perception and tissue handling showed a trend toward improvement, the curves were not statistically significant.

**Table 2** Descriptive statistics—assessment of laparoscopic cases

|                          | Surgical records |       | Unique fellows<br>$n^a$ | No. previous complex cases |        |       | No. previous cases of same type |        |       |
|--------------------------|------------------|-------|-------------------------|----------------------------|--------|-------|---------------------------------|--------|-------|
|                          | $n$              | (%)   |                         | Mean                       | Median | Range | Mean                            | Median | Range |
| Total                    | 98               | 100   | 31                      | 79                         | 67.5   |       | 21.3                            | 9.5    |       |
| LapBand placement        | 17               | 17.35 | 10                      | 58.6                       | 50     | 6–177 | 6.8                             | 6      | 0–22  |
| Nissen fundoplication    | 27               | 27.55 | 15                      | 119                        | 100    | 0–360 | 15.1                            | 8      | 0–60  |
| Roux-en-Y gastric bypass | 54               | 55.1  | 20                      | 65.4                       | 67.5   | 0–223 | 29.0                            | 13     | 0–160 |

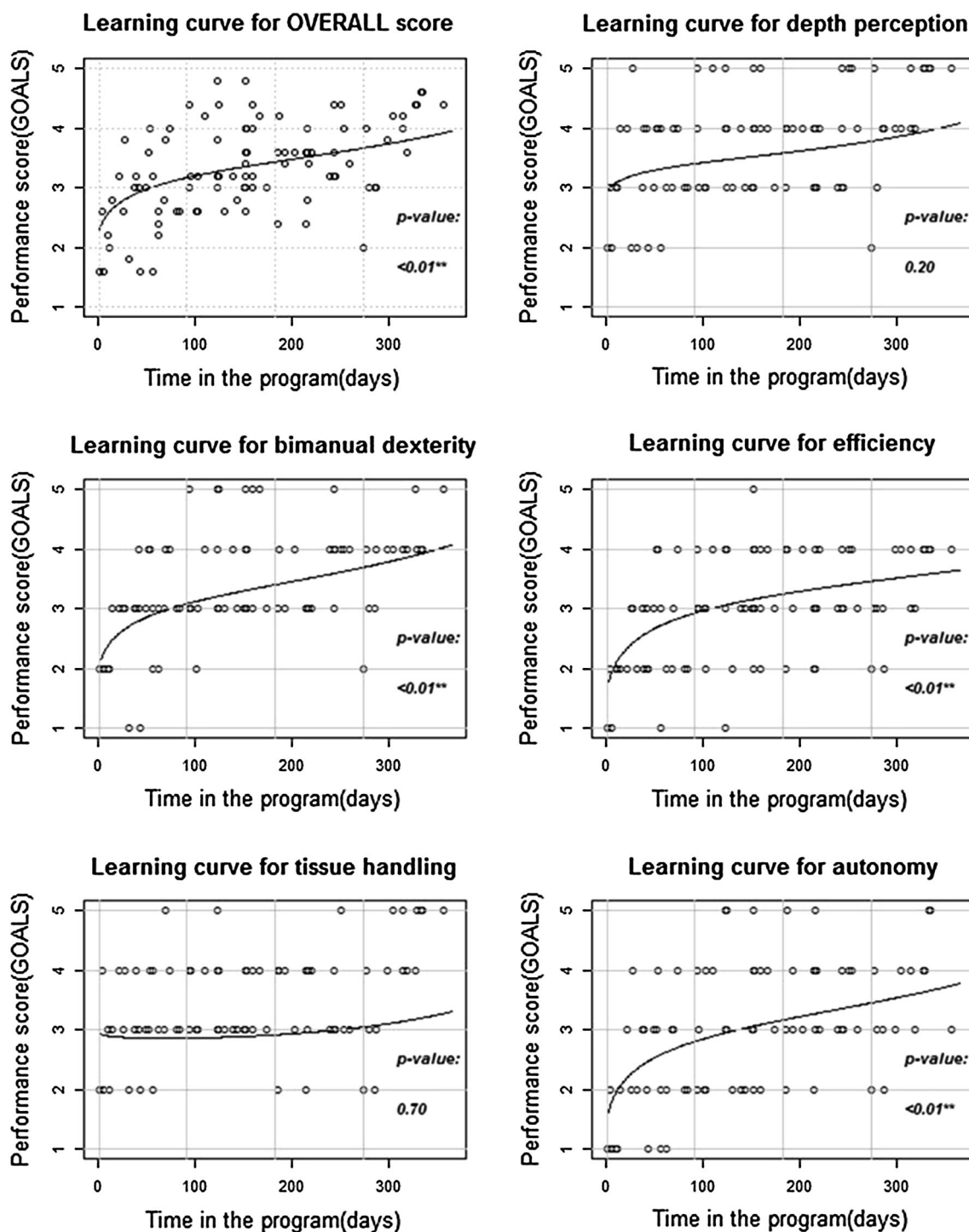
<sup>a</sup> Unique fellows total more than 31 because some fellows performed more than 1 procedure each



**Fig. 1** Mean score of overall and five performance domains across four quarters

**Table 3** Modeling results for four predictive factors

|  | Separate models      |       | <i>P</i> value | Full model           |       | <i>P</i> value | Final Model          |  |
|--|----------------------|-------|----------------|----------------------|-------|----------------|----------------------|--|
|  | Coefficient estimate |       |                | Coefficient estimate |       |                | Coefficient estimate |  |
|  | Intercept            | Slope |                |                      |       |                |                      |  |
| Total number of previous complex cases | 3.3                  | 0.006 | 0.01           | 0.003                | 0.01  | 0.003          | <0.01                |  |
| Number of previous cases of same type  | 3.0                  | 0.004 | <0.01          | 0.003                | 0.10  | –              |                      |  |
| Difficulty level                       | 3.0                  | 0.133 | 0.02           | 0.08                 | 0.09  | 0.106          | 0.02                 |  |
| Time in the program (log)              | 1.4                  | 0.427 | <0.01          | 0.32                 | <0.01 | 0.312          | <0.01                |  |



**Fig. 2** Learning curves for overall performance and five performance domains (GOALS)

## Discussion

Competency-based education is currently being introduced into all levels of surgical training in the U.S. Fellowship training in complex surgery is continually evolving, not only to keep up with the advances in patient care and the

adoption of new technologies, but also to meet the needs of trainees. Valid and reliable performance assessment tools are essential to ensure that competencies are acquired. Some tools have been proven to be valid measures of technical performance, although none has been adopted for generalized use for technical skills assessments.

In addition to providing an assessment with which to provide feedback, data from assessments can be analyzed with other variables to identify factors that result in improved performance. In this project, we developed several analytic models to identify which model would best identify these predictors. In our Final Model, we were able to identify and confirm that three separate factors all independently impacted the performance scores. In future studies with larger data sets, we will further test these models. This study provides evidence that we can use these models to identify factors that impact the performance scores of trainees.

By comparing performance scores of fellows at the beginning of the fellowship year with their scores at the end of the fellowship year, we determined that GOALS is able to differentiate novice fellows from graduating fellows and, therefore, we have established construct validity for the tool. These results confirm that GOALS may be reliably used to provide feedback to fellows by program directors and other faculty. Formative feedback of this type has the potential to allow program directors to customize training for each fellow to meet her/his specific needs for technical skills training. Ultimately, this type of tool could be used for summative assessment to inform critical decisions such as graduation, certification, and credentialing.

Learning curves graphically display the change in performance over time during a period of learning. When the learning curve reaches a plateau, learning has been completed. A trainee will have achieved proficiency in performing an operation if her/his scores plateau at a high level of performance that would be consistent with proficiency. If the scores plateau at lower levels of performance, either the trainee or the training methods may have deficiencies. However, for surgical educators to make summative assessments with tools such as GOALS, they must be validated.

The limitations of this study are the small number of fellows who were assessed, the small number of assessments for each fellow, and the small number of fellows for whom there were multiple assessments for the same type of case throughout the fellowship year. Despite that, the scores for the fellows represented in the dataset tended to improve throughout the fellowship year, strongly suggesting that, as a group, the fellows performed significantly better as the year passed.

Although the performance scores improved through the course of the year overall and for all domains, that improvement did not reach statistical significance either in the domain of depth perception or of tissue handling. One possible explanation is that the number of fellows may have been too small to provide sufficient power. Another possible explanation is that the fellows had already acquired significant skill in these domains before becoming

a fellow. This possibility is supported by the fact that the performance scores in these two domains at the beginning of the fellowship were higher than the performance scores in the other domains. All fellows had previously completed a general surgery residency and had a large operative experience with laparoscopic cholecystectomy. They may have acquired the skills to effectively manage the lack of depth perception and effectively handle tissue while performing those operations.

The ultimate value of assessment tools, such as GOALS, is that they will enable training to technical proficiency. Training surgeons to proficiency will not only enable surgical training programs to graduate better-trained surgeons but also enable surgical educators to document that each graduating surgeon is well trained.

## Conclusions

This study has documented that GOALS is able to differentiate novice fellows from graduating fellows and, therefore, the study establishes construct validity for GOALS as an assessment tool for technical performance during complex laparoscopic gastrointestinal operations. The analytic models developed and tested in this study may now be studied on larger, more complete data sets as they become available. These future studies will better define optimal use of both the GOALS tool and the analytic models.

**Acknowledgments** The study was funded by AHRQ Grant No. R03 HS021549, Determining a Learning Curve for Complex Laparoscopic Gastrointestinal Surgery.

**Disclosure** Nancy J Hogle, Yi Liu, R Todd Ogden, and Dennis L Fowler have no conflicts of interest or financial ties to disclose.

## References

1. Barnes RW (1987) Surgical handicraft: teaching and learning surgical skills. *Am J Surg* 153(5):422–427
2. Grober ED, Jewett MA (2006) The concept and trajectory of “operative competence” in surgical training. *Can J Surg* 49(4): 238–240
3. Moulton CA, Dubrowski A, Macrae H, Graham B, Grober E, Reznick R (2006) Teaching surgical skills: what kind of practice makes perfect? A randomized, controlled trial. *Ann Surg* 244(3):400–409
4. Moorthy K, Munz Y, Sarker SK, Darzi A (2003) Objective assessment of technical skills in surgery. *BMJ* 327(7422):1032–1037
5. Buth J, Harris PL, Maurer PC, Nachbur B, van Urk H (2000) Harmonization of vascular surgical training in Europe. A task for the European Board of Vascular Surgery (EBVS). *Cardiovasc Surg* 8(2):98–103
6. Darzi A, Datta V, Mackay S (2001) The challenge of objective assessment of surgical skill. *Am J Surg* 181(6):484–486

7. Darzi A, Smith S, Taffinder N (1999) Assessing operative skill. Needs to become more objective. *BMJ* 318(7188):887–888
8. Mackay S, Datta V, Chang A, Shah J, Kneebone R, Darzi A (2003) Multiple objective measures of skill (MOMS): a new approach to the assessment of technical ability in surgical trainees. *Ann Surg* 238(2):291–300
9. MacRae H, Regehr G, Leadbetter W, Reznick RK (2000) A comprehensive examination for senior surgical residents. *Am J Surg* 179(3):190–193
10. MacRae HM, Cohen R, Regehr G, Reznick R, Burnstein M (1997) A new assessment tool: the patient assessment and management examination. *Surgery* 122(2):335–343 (discussion 343–334)
11. Martin JA, Regehr G, Reznick R et al (1997) Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 84(2):273–278
12. Pandey V, Jackson M, Bicknell C, Yapanis M, Darzi A, Wolfe J (2003) Assessing competence in vascular surgery: a pilot study. *Br J Surg* 90(S1):113
13. Reznick RK (1993) Teaching and testing technical skills. *Am J Surg* 165(3):358–361
14. Satava RM, Cuschieri A, Hamdorf J (2003) Metrics for objective assessment. *Surg Endosc* 17(2):220–226
15. Accreditation Council of Graduate Medical Education. <http://www.acgme-2010standards.org/approved-standards.html>. Accessed 29 Sept 2011
16. Vassiliou MC, Feldman LS, Andrew CG et al (2005) A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg* 190(1):107–113
17. Gumbs AA, Hogle NJ, Fowler DL (2007) Evaluation of resident laparoscopic performance using Global Operative Assessment of Laparoscopic Skills. *J Am Coll Surg* 204(2):308–313
18. Chang L, Hogle NJ, Moore BB et al (2007) Reliable assessment of laparoscopic performance in the operating room using videotape analysis. *Surg Innov* 14(2):122–126
19. Kurashima Y, Feldman LS, Al-Sabah S, Kaneva PA, Fried GM, Vassiliou MC (2011) A tool for training and evaluation of laparoscopic inguinal hernia repair: the Global Operative Assessment of Laparoscopic Skills-Groin Hernia (GOALS-GH). *Am J Surg* 201(1):54–61
20. Ghaderi I, Vaillancourt M, Sroka G et al (2011) Evaluation of surgical performance during laparoscopic incisional hernia repair: a multicenter study. *Surg Endosc* 25(8):2555–2563
21. Vaillancourt M, Ghaderi I, Kaneva P et al (2011) GOALS-incisional hernia: a valid assessment of simulated laparoscopic incisional hernia repair. *Surg Innov* 18(1):48–54
22. Schwarz GE (1978) Estimating the dimension of a model. *Ann Stat* 6(2):461–464