There is no objective scale for assessment of operative skill in laparoscopic gastric bypass (LGBP). The objective of this study was to develop and demonstrate feasibility of use, validity, and reliability of a Bariatric Objective Structured Assessment of Technical Skill (BOSATS) scale.

The BOSATS scale was developed using a hierarchical task analysis (HTA), a Delphi questionnaire, and a panel of international experts in bariatric surgery. The feasibility of use, reliability, and validity of the developed scale were demonstrated by reviewing 52 prospectively collected video recordings of LGBP performed by novice and experienced surgeons.

A total of 214 discrete steps were identified in HTA. A total of 12 and 17 panel members completed the first and second round of the Delphi questionnaire, respectively. Consensus among the panel was achieved after the second round (Cronbach’s alpha = 0.85). The BOSATS scale demonstrated high inter-rater (intraclass correlation coefficient [ICC] = 0.954; p < 0.001) and test-retest reliability (ICC = 0.99; p < 0.001). Significant differences between BOSATS scores of experienced and novice surgeon groups were noted for the creation of jejunojejunostomy (JJ), gastric pouch, linear stapled gastrojejunostomy (GJ), circular stapled GJ, and hand-sewn GJ. Moderate to high correlations between BOSATS scale and Objective Structured Assessment of Technical Skills Global Rating Scale (OSATS GRS) were seen for JJ (rho = 0.59; p = 0.001), gastric pouch (rho = 0.48; p = 0.0004), linear stapled GJ (rho = 0.70; p = 0.0001), and hand-sewn GJ (rho = 0.96; p < 0.0001).

The BOSATS scale is a feasible to use, reliable, and valid instrument for objective assessment of operative performance in LGBP. Implementation of this scale is expected to facilitate deliberate practice and provide a means for future certification in bariatric surgery. (J Am Coll Surg 2013;216:955–965. © 2013 by the American College of Surgeons)

Over the past 2 decades, laparoscopic Roux-en-Y gastric bypass (LGBP) has gained increasing popularity in North America as the procedure of choice for morbid obesity. More than 50,000 LGBP operations were performed in the United States in 2006. Achieving proficiency in this advanced laparoscopic procedure remains a challenge for both surgical residents and practicing surgeons. With a learning curve of 50 to 100 cases, surgeons’ technical
Operative Assessment of Laparoscopic Skills (GOALS), operation that need the most improvement. Unfortunately, global rating scales do not provide specific information on specific parts of an operation, etc. Such scales, unlike procedure-specific scales, are designed to provide trainees with a global assessment of their operative performance (respect for tissue, time and motion, etc) performed on bench top models. The GOALS and modified OSATS GRS scales were developed and validated for use in laparoscopic surgery. Examples of such scales include the Objective Structured Assessment of Technical Skills Global Rating Scale (OSATS GRS), modified OSATS GRS, Global Operative Assessment of Laparoscopic Skills (GOALS), as well as procedure-specific assessment scales for laparoscopic Nissen fundoplication and laparoscopic right and sigmoid colectomies. None of these scales has been developed or validated for use in LGBP.

The OSATS GRS, developed by Reznick and colleagues, is probably the most widely used form of objective assessment of technical skill in surgery. It was developed and validated for use with a specific set of surgical tasks (bowel anastomosis, vascular anastomosis, etc) performed on bench top models. The GOALS and modified OSATS GRS scales were developed as alternatives to OSATS GRS for laparoscopic surgery; however, both scales were validated only for use in laparoscopic cholecystectomy. The OSATS GRS, modified OSATS GRS, and GOALS are examples of global rating scales. Such scales, unlike procedure-specific scales, are designed to provide trainees with a global assessment of their operative performance (respect for tissue, time and motion, etc). Unfortunately, global rating scales do not provide a trainee with any information on specific parts of an operation that need the most improvement.

At present, there is no objective scale to assess operative performance in LGBP. Aggarwal and colleagues were the only group that attempted to address this deficiency by designing and validating a global and procedure-specific assessment scale for the operative step of laparoscopic jejunojejunostomy (JJ). Albeit an excellent start, that scale was designed to assess only 1 component of the LGBP operation. Furthermore, it is of limited benefit in live surgery because it was validated only for use on cadaveric porcine models.

An ideal scale for assessment of surgical proficiency in LGBP should be comprehensive, feasible to use, internationally relevant, reliable, and valid for live and recorded surgery. It should include all common surgical approaches to the LGBP, enabling its use in different institutions and enabling comparability between surgeons. Use of such a scale would not only allow for an objective evaluation of a surgeon’s technical proficiency in LGBP, but would also provide a basis for constructive feedback and deliberate practice.

This study had 3 objectives. The first objective was to deconstruct the LGBP procedure into its component steps. The second objective was to use the Delphi method and an international panel of experts in bariatric surgery to select the steps of LGBP that were considered important for inclusion into a Bariatric Objective Structured Assessment of Technical Skill (BOSATS) scale. The third objective was to demonstrate feasibility of use, inter-rater and test-retest reliability, as well as construct and concurrent validity for the developed scale.

**METHODS**

**Hierarchical task analysis**

A hierarchical task analysis (HTA) of LGBP was carried out to deconstruct the operation into its component steps. A similar approach has been used in the past to define steps for some basic and intermediate laparoscopic procedures. Ten full-length video recordings of LGBP performed using various surgical approaches (linear stapled JJ, linear stapled gastrojejunostomy [GJ], circular stapled GJ, hand-sewn GJ, antecolic antegastric and retrocolic retrogastric Roux limb placement) were reviewed independently by 2 experienced laparoscopic surgeons (RA and TG). Each reviewer listed all consecutive steps required for completion of each operation. An in-person meeting was organized between the reviewers to resolve any disparities in regard to the generated steps. The list of steps for completion of each operation was organized between the reviewers to resolve any disparities in regard to the generated steps. The list of steps for completion of each operation was organized between the reviewers to resolve any disparities in regard to the generated steps. The list of steps for completion of each operation was organized between the reviewers to resolve any disparities in regard to the generated steps. The list of steps for completion of each operation was organized between the reviewers to resolve any disparities in regard to the generated steps.

**Selection of steps for inclusion into the BOSATS scale**

**Study design**

The Delphi method and an online questionnaire were used to achieve consensus among a panel of international
experts in bariatric surgery on the steps of LGBP that were considered of greatest importance for inclusion into the BOSATS scale. The objective for this part of the study was to select steps from the HTA that were important for assessing surgical proficiency in LGBP—performed using various surgical approaches—and not to achieve expert consensus on the best surgical approach. The Delphi method was used as a means for obtaining opinions from a panel of experts in a systematic manner.\textsuperscript{14-17} Responses of the panel were collated, analyzed, and presented back in an iterative fashion until consensus was achieved. Research ethics boards of St Michael’s Hospital and University of Toronto, Toronto, Canada approved the study.

**Selection of the expert panel**
Members of the expert panel were selected based on their roles as international opinion leaders in the field of advanced laparoscopic surgery, as evidenced by their roles within North American, South American, European, and Australasian surgical societies. All members of the expert panel were also members of the Advanced Training in Laparoscopic Abdominal Surgery (ATLAS) group. Participation in the study was voluntary and informed consent was implied if an individual agreed to participate.

**Administration of the questionnaire**
The list of steps for LGBP generated from HTA was presented to each panel member via an online questionnaire (www.surveymonkey.com). Each panel member was asked to rate the operative steps on a Likert-type scale from 1 (“strongly disagree”) to 5 (“strongly agree”) with respect to the degree of relevance for inclusion into the BOSATS scale. Panel members were given an opportunity to comment on their selections and to clarify their ratings. Two questionnaire rounds were administered and a total of 3 email reminders for questionnaire completion were sent during each round.

**Determination of consensus**
Cronbach’s alpha was chosen as the statistical index to quantify a measure of consensus among panel members.\textsuperscript{17} Bland and Altman\textsuperscript{18} and Shrout and Fleiss\textsuperscript{19} suggested that Cronbach’s alpha should be above 0.90 for a diagnostic scale to be useful in clinical practice. For educational interventions, alpha is often set at 0.70.\textsuperscript{20} For this study, alpha $\geq 0.80$ was chosen to represent an acceptable measure of consensus. If consensus was not achieved, group means and standard deviations for each step were calculated, steps were added or modified based on the comments of the panel members, and the questionnaire was resubmitted for rating. This process continued in an iterative fashion until consensus was achieved.

**Selection of steps for inclusion into the BOSATS scale**
Once consensus among the panel members was achieved (alpha $\geq 0.80$), steps that were rated as 4 (“agree”) or 5 (“strongly agree”) by more than 80% of panel members were selected for inclusion into the BOSATS scale. Anchoring descriptors for technical proficiency for each step were assigned to a numerical scoring scale ranging from 1 to 5.

**Statistical analysis**
Descriptive statistics were calculated for each step. Cronbach’s alpha was calculated as a measure of consensus among panel members. Missing data points, which resulted from submissions of incomplete questionnaires, were dealt with in 3 ways: the missing data point was replaced with a mean for that step; the missing data point was replaced with a 3 (“neutral”) based on the assumption that because the step was omitted, the panel member did not feel strongly about including or excluding that step; and the missing data point was replaced with the mode for that step. Cronbach’s alpha was calculated using each of the above methods.

**Determination of feasibility of use, validity and reliability of the BOSATS scale**

**Study design**
This part of the study used a prospective single-blinded observational study design. Video recordings of LGBP performed by senior surgical residents (postgraduate years 3 to 5), minimally invasive surgery fellows, and experienced laparoscopic bariatric surgeons at 3 academic teaching hospitals in Ontario, Canada were prospectively collected. Only the laparoscopic intra-abdominal camera view was recorded. The operative approaches to LGBP were: creation of JJ (linear stapled anastomosis with hand-sewn closure of common enterotomy), positioning of the Roux limb (antecolic antegastric approach), and creation of GJ (circular stapled anastomosis with transoral anvil placement, linear stapled and hand-sewn anastomoses).

**Participant selection**
Study participants were selected based on case-volume criteria. Novice surgeons were defined as individuals who had acted as a primary surgeon in fewer than 10 LGBP operations in its entirety. Minimally invasive surgical fellows who had performed fewer than 10 LGBP operations were considered novices. Experienced
surgeons were defined as individuals who had performed more than 100 LGBP procedures as a primary surgeon. Experienced surgeons were expected to be beyond their initial learning curve. All participants were required to sign an informed consent before participating in this study.

**Video assessments**

Two trained and independent raters (BZ and EB), blinded to the identity of the operating surgeon, reviewed LGBP video recordings and scored them using the BOSATS scale. If a step on the BOSATS scale was not performed or visualized on the video, the raters were instructed to rate it as “not applicable.” In addition, each discrete component of the operation (creation of JJ, creation of gastric pouch, and creation of GJ) was also evaluated using the OSATS GRS scale. This approach was chosen to account for the common practice of novice surgeons learning LGBP in steps (completion of JJ, completion of GJ, etc) before completing the operation in its entirety. Furthermore, evaluating the entire procedure with the OSATS GRS would have been methodically incorrect because procedures were often shared between experienced and novice surgeons.

**Sample size**

A power calculation was performed a priori based on earlier work with the GOALS scale. The minimum relevant difference between novice and experienced groups in that study was 6.4 points. Using a standard deviation of 4.5, power of 0.8, and alpha of 0.05, the minimum required number of videos for each group was 8.

**Feasibility of use**

All minimally invasive operating rooms at the academic hospitals involved in this study were equipped with video-recording equipment. It is the standard of practice at these hospitals to record all laparoscopic cases. As a result, there were no associated costs with obtaining the video recordings. In regard to the use of the BOSATS scale, raters were asked to record the average time required to score an LGBP video recording. Fast-forwarding of video recordings was permitted.

**Reliability**

Inter-rater reliability for the BOSATS scale was calculated by correlating total and component scores between 12 independent raters using the intraclass correlation coefficient (ICC, 2-way mixed-effects model, absolute agreement). Test-retest reliability for the BOSATS scale was calculated by correlating the total score for case number 1 assessed in its entirety at 2 different time points using ICC.

**Validity**

Differences between novice and experienced surgeon group scores were analyzed to provide evidence of construct validity for each of the following components: creation of JJ, creation of gastric pouch, and creation of GJ. Surgeons were assigned to a novice or an experienced group based on previously defined case-volume criteria. In view of the small sample size, data were treated as nonparametric. Scores of 2 raters for each component were pooled together and the complete data set was analyzed using a Mann-Whitney U test to assess differences between groups. Bonferroni correction was applied to account for a potential for repeated measures in view of pooling rater scores. Statistical significance was set to p < 0.025. Group scores were reported as median (interquartile range).

Data were then reanalyzed using the following definitions for novice and experienced surgeons: surgeons who scored <28 on the OSATS GRS scale were assigned to the novice group, and surgeons that scored ≥28 were assigned to the experienced group. This method of grouping avoided the inherent subjectivity associated with classification of surgeons based on their case volumes.

Concurrent validity for creation of JJ, gastric pouch, and creation of GJ components was calculated by correlating BOSATS scores with OSATS GRS scores using Spearman’s correlation coefficient.

**Statistical analysis**

All statistical analyses were performed using STATA version 12.0.

**RESULTS**

**Hierarchical task analysis**

A total of 14 full-length video recordings of LGBP, performed using the following common approaches, were reviewed: creation of JJ anastomosis (linear stapled anastomosis with hand-sewn or stapled closure of common enterotomy), placement of the Roux limb (antecolic antegastric, retrocolic retrogastric, retrocolic antegastric), creation of GJ anastomosis (circular stapled with transgastric and transoral anvil placement, linear stapled, hand-sewn). A total of 214 discrete steps were identified for all examined approaches. A complete list of steps is available from the authors on request.

**Selection of steps for inclusion into the BOSATS scale**

Twelve international experts in laparoscopic bariatric surgery from 5 countries were invited to complete the first
round of the online Delphi questionnaire. All 12 individuals (100%) completed the questionnaire (Table 1). Cronbach’s alpha for the entire questionnaire (containing all steps from HTA) was 0.64 when missing data points were replaced by mean, 0.55 when missing data points were replaced by 3 “neutral,” and 0.62 when missing data points were replaced by mode. Cronbach’s alpha for individual components of the questionnaire ranged from 0.25 to 1.00 (Table 2).

A total of 19 international experts in laparoscopic bariatric surgery from 7 countries were invited to complete the second round of the online Delphi questionnaire. Seventeen participants (89%) from 6 countries completed the survey (Table 1). Cronbach’s alpha values for the entire questionnaire were 0.85, 0.83, and 0.84 when missing data points were replaced by mean, 3 “neutral,” and mode, respectively. Cronbach’s alpha for individual components of questionnaire ranged from 0.15 to 0.97 (Table 2). Once consensus among panel members was achieved, a total of 99 of 214 (46%) steps from the questionnaire were selected for inclusion into the BOSATS scale (Appendix 1, online only).

### Determination of feasibility of use, validity and reliability of the BOSATS scale

#### Reliability

A total of 52 video recordings of LGBP were reviewed and scored by 2 independent, trained, and blinded raters using the BOSATS scale. The inter-rater reliability coefficient for the BOSATS scale was ICC = 0.954 (p < 0.001). The inter-rater reliability coefficients for individual components of the BOSATS scale ranged from ICC = 0.481 to 0.924 (p < 0.05; Table 3). The test-retest reliability coefficients for the BOSATS scale were ICC = 0.992 (p < 0.001) and ICC = 0.985 (p < 0.001) for rater 1 and rater 2, respectively.

#### Feasibility of use

The average time to evaluate a 120-minute video recording of an uncomplicated LGBP operation using the BOSATS scale was 30 to 40 minutes.

#### Construct validity

Using case-volume criteria for group assignments, significant differences between novice and experienced surgeon groups were seen for the following components of the

### Table 1. Composition of the International Expert Panel for the Online Delphi Questionnaire

<table>
<thead>
<tr>
<th>Location</th>
<th>Participants in first round, n</th>
<th>Participants in the second round, n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Contacted</td>
<td>Responded</td>
</tr>
<tr>
<td>Canada</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>USA</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>UK</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Chile</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Australia</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Denmark</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Sweden</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>12</td>
</tr>
</tbody>
</table>

Response rate: 100% 89%

### Table 2. Agreement among Panel Members for First and Second Round of Online Delphi Questionnaire

<table>
<thead>
<tr>
<th>Component of the questionnaire</th>
<th>First round</th>
<th>Second round</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n, alpha</td>
<td>n, alpha</td>
</tr>
<tr>
<td>Entire questionnaire (all components included)</td>
<td>12, 0.55</td>
<td>17, 0.83</td>
</tr>
<tr>
<td>Patient positioning</td>
<td>12, 0.89</td>
<td>17, 0.92</td>
</tr>
<tr>
<td>Abdominal access</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Veress needle technique</td>
<td>11, 0.68</td>
<td>8, 0.61</td>
</tr>
<tr>
<td>Optical trocar without prior insufflation</td>
<td>12, 0.62</td>
<td>9, 0.64</td>
</tr>
<tr>
<td>Diagnostic laparoscopy and port placement</td>
<td>12, 0.74</td>
<td>16, 0.96</td>
</tr>
<tr>
<td>Placement of liver retractor</td>
<td>12, 0.90</td>
<td>17, 0.73</td>
</tr>
<tr>
<td>Creation of the Roux limb</td>
<td>12, 0.56</td>
<td>16, 0.86</td>
</tr>
</tbody>
</table>

#### Jejunoojejunostomy

- Linear stapled with hand-sewn or stapled closure of common enterotomy | 12, 0.81 | 17, 0.90 |
- Trippele stapling technique | 11, 0.61 | 1 — |
- Dissection of the phreno-esophageal ligament | 12, 0.89 | 17, 0.83 |
- Creation of gastric pouch | 12, 0.64 | 17, 0.85 |
- Roux limb positioning |                        |              |
- Antecolic, antegastric | 12, 0.75 | 15, 0.92 |
- Retecolic, retrogastric | 9, 1.00 | 2 — |
- Retecolic, antegastric | 10, 0.25 | 2 — |
- Creation of gastrojejunostomy |                        |              |
- Linear stapled anastomosis with hand-sewn closure of common enterotomy | 9, 0.81 | 10, 0.90 |
- Circular stapled anastomosis | 11, 0.63 | 7, 0.57 |
- Hand-sewn anastomosis | 11, 0.58 | 7, 0.24 |
- Closure of mesenteric defects |                        |              |
- Jejunoojejunostomy defect | 12, 0.85 | 17, 0.81 |
- Petersen’s defect | 11, 0.49 | 11, 0.29 |
- Transverse mesocolon defect | 9, 0.74 | 3 — |
- Gastrojejunostomy leak test |                        |              |
- Blue dye | 12, 0.36 | 9 — |
- Air insufflation | 10, 1.00 | 12, 0.15 |
- Removal of ports and retractors | 12, 0.90 | 17, 0.65 |
- Wound closure | 12, 0.97 | 17, 0.97 |

Missing data points were replaced by 3 “neutral” for analysis. Data reported as n, number of participants responding; alpha denotes Cronbach’s alpha.
BOSATS scale: creation of JJ, linear stapled GJ, circular stapled GJ, and hand-sewn GJ (Table 5; Fig. 2). All circular stapled GJ videos were grouped into the experienced surgeon group using OSATS GRS criteria, prohibiting calculation of construct validity for this component of the BOSATS scale.

**Concurrent validity**

Moderate to strong correlations between BOSATS scores and OSATS GRS scores were noted for the following components of LGBP: creation of JJ, creation of gastric pouch, linear stapled GJ, and hand-sewn GJ (Table 6; Fig. 3). A moderate, but not statistically significant, correlation between BOSATS scores and OSATS GRS scores was seen for circular stapled GJ (Table 6).

**DISCUSSION**

A scale for objective assessment of operative performance in LGBP (BOSATS) was developed and validated using a systematic and evidence-based approach. First, a hierarchical task analysis of various surgical approaches to LGBP was conducted to generate a list of all the steps required to complete this procedure. Second, this list was presented in the form of an online Delphi questionnaire to a panel of international experts in bariatric surgery, with a goal of identifying steps that were essential for inclusion into the BOSATS scale. Anchoring descriptors for levels of surgical proficiency were then assigned to each step in the scale. Last, feasibility of use, reliability, and validity of the BOSATS scale were established using blinded and independent review of prospectively recorded LGBP procedures.

The BOSATS scale has several strengths. It is the only objective procedure-specific assessment scale specifically developed and validated for use in LGBP. Items for this scale were selected using a Delphi method and consensus of an international panel of experts in bariatric surgery, providing evidence for content validity. The BOSATS scale has several strengths. It is the only objective procedure-specific assessment scale specifically developed and validated for use in LGBP. Items for this scale were selected using a Delphi method and consensus of an international panel of experts in bariatric surgery, providing evidence for content validity.

**Table 3.** Inter-Rater Reliability Coefficients for the Complete Bariatric Objective Structured Assessment of Technical Skill Scale and its Individual Components

<table>
<thead>
<tr>
<th>Component of BOSATS scale</th>
<th>n</th>
<th>ICC</th>
<th>95% CI</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total score</td>
<td>52</td>
<td>0.954</td>
<td>0.922–0.974</td>
<td>0.000</td>
</tr>
<tr>
<td>Access and port insertion</td>
<td>23</td>
<td>0.924</td>
<td>0.883–0.967</td>
<td>0.000</td>
</tr>
<tr>
<td>Placement of liver retractor</td>
<td>34</td>
<td>0.721</td>
<td>0.508–0.850</td>
<td>0.000</td>
</tr>
<tr>
<td>Creation of the Roux limb</td>
<td>27</td>
<td>0.481</td>
<td>0.323–0.723</td>
<td>0.006</td>
</tr>
<tr>
<td>Creation of JJ anastomosis</td>
<td>29</td>
<td>0.813</td>
<td>0.640–0.908</td>
<td>0.000</td>
</tr>
<tr>
<td>Dissection of the phrenoesophageal ligament</td>
<td>34</td>
<td>0.485</td>
<td>0.189–0.703</td>
<td>0.001</td>
</tr>
<tr>
<td>Creation of gastric pouch</td>
<td>50</td>
<td>0.717</td>
<td>0.549–0.829</td>
<td>0.000</td>
</tr>
<tr>
<td>Positioning of the Roux limb</td>
<td>46</td>
<td>0.830</td>
<td>0.713–0.902</td>
<td>0.000</td>
</tr>
<tr>
<td>Creation of GJ anastomosis (linear stapler)</td>
<td>26</td>
<td>0.647</td>
<td>0.348–0.825</td>
<td>0.000</td>
</tr>
<tr>
<td>Creation of GJ anastomosis (circular stapler)</td>
<td>10</td>
<td>0.639</td>
<td>0.113–0.893</td>
<td>0.013</td>
</tr>
<tr>
<td>Creation of GJ anastomosis (hand-sewn)</td>
<td>15</td>
<td>0.831</td>
<td>0.569–0.940</td>
<td>0.000</td>
</tr>
<tr>
<td>Closure of JJ mesenteric defect</td>
<td>21</td>
<td>0.704</td>
<td>0.401–0.868</td>
<td>0.000</td>
</tr>
<tr>
<td>Testing of GJ anastomosis (air insufflation)</td>
<td>27</td>
<td>0.877</td>
<td>0.750–0.942</td>
<td>0.000</td>
</tr>
<tr>
<td>Removal of liver retractor</td>
<td>41</td>
<td>0.726</td>
<td>0.540–0.844</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Data reported as n, number of cases analyzed.

BOSATS, Bariatric Objective Structured Assessment of Technical Skill; GJ, gastrojejunostomy; ICC, intraclass correlation coefficient; JJ, jejunojejunostomy.

**Table 4.** Comparison of Scores for Different Components of Gastric Bypass as Assessed by Bariatric Objective Structured Assessment of Technical Skill Scale Using Case-Volume Criteria for Definition of Experienced and Novice Surgeons

<table>
<thead>
<tr>
<th>Component of BOSATS scale</th>
<th>Observations, n</th>
<th>BOSATS score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Novice</td>
<td>Experienced</td>
</tr>
<tr>
<td>Jejunojejunostomy</td>
<td>22</td>
<td>30</td>
</tr>
<tr>
<td>Gastric pouch</td>
<td>32</td>
<td>68</td>
</tr>
<tr>
<td>Linear stapled gastrojejunostomy</td>
<td>14</td>
<td>38</td>
</tr>
<tr>
<td>Circular stapled gastrojejunostomy</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Hand-sewn gastrojejunostomy</td>
<td>14</td>
<td>16</td>
</tr>
</tbody>
</table>

Scores reported as median (interquartile range).

*Significant differences between groups (p < 0.025).

BOSATS, Bariatric Objective Structured Assessment of Technical Skill Scale.
scale was intentionally designed to address multiple approaches to LGBP, increasing its transferability between surgeons and institutions.

Procedure-specific assessment scales, such as the scale developed in this study, are distinctly different from task-specific checklists and global rating scales. Task-specific checklists, such as those for laparoscopic sigmoid colectomy, Nissen fundoplication, cholecystectomy, and inguinal hernia repair, deconstruct the operation into distinct tasks and mandate a surgeon to perform those tasks in a predefined order. Failure to follow this predefined order often results in a low score on the checklist. Checklists have been shown to provide trainees with structured formative feedback; however, their rigid nature has led to the finding of poor validity and reliability when used by proficient surgeons. Global rating scales, on the other hand, look at global parameters of operative performance. They tend to be more reliable than task-specific checklists; however, global rating scales are not very useful for providing specific feedback to the learner.

Procedure-specific assessment scales form the middle ground between task-specific checklists and global rating scales. They are less rigid than task-specific checklists, while still providing specific formative feedback to the learner.

The Delphi method used in this study had several strengths. It offered the opportunity to conduct the questionnaire online, thereby improving feasibility and lowering costs. It allowed for recruitment of participants from various geographic locations, and the anonymous

**Table 5.** Comparison of Scores for Different Components of Gastric Bypass as Assessed by Bariatric Objective Structured Assessment of Technical Skill Scale Using Objective Structured Assessment of Technical Skills Global Rating Scale Criteria for Definition of Experienced and Novice Surgeons

<table>
<thead>
<tr>
<th>Component</th>
<th>Observations, n</th>
<th>Novice</th>
<th>Experienced</th>
<th>p Value</th>
<th>Maximum possible score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jejunojejunostomy</td>
<td>22</td>
<td>32</td>
<td>26.5 (24–23)*</td>
<td>0.007*</td>
<td>35*</td>
</tr>
<tr>
<td>Gastric pouch</td>
<td>20</td>
<td>78</td>
<td>34.5 (29.5–38)*</td>
<td>0.008*</td>
<td>50*</td>
</tr>
<tr>
<td>Linear stapled gastrojejunostomy</td>
<td>16</td>
<td>34</td>
<td>27 (23–29.5)*</td>
<td>&lt;0.0001</td>
<td>35*</td>
</tr>
<tr>
<td>Circular stapled gastrojejunostomy</td>
<td>0</td>
<td>20</td>
<td>—</td>
<td>—</td>
<td>50</td>
</tr>
<tr>
<td>Hand-sewn gastrojejunostomy</td>
<td>14</td>
<td>14</td>
<td>15 (13–16)*</td>
<td>0.0001*</td>
<td>25*</td>
</tr>
</tbody>
</table>

Experience surgeons defined as Objective Structured Assessment of Technical Skills ≥ 28; novice surgeons as < 28. Scores are reported as median (interquartile range).

*Significant differences between groups (p < 0.025).

BOSATS, Bariatric Objective Structured Assessment of Technical Skill.
nature of the Delphi method made it challenging for a single influential participant to have a disproportionate impact on the outcome of the questionnaire. Response rates for the first and second rounds of the questionnaire were higher than the response rates commonly reported in health professional literature, which was likely a reflection of participants’ interest in the presented topic. Despite excellent response rates, however, 2 rounds of the online Delphi questionnaire were required to achieve consensus among the panel members. This lack of initial consensus speaks to the diversity of opinions among international experts on the steps of LGBP that were of importance for inclusion into the BOSATS scale. Studies by Palter and colleagues, which used similar methodologies to develop an assessment scale for laparoscopic colorectal surgery, reported similar trends.

The BOSATS scale demonstrated excellent inter-rater and test-retest reliability. Inter-rater reliability is a psychometric property of a scale that refers to the extent to which ratings of the same performance by different observers are similar. The BOSATS scale demonstrated very high inter-rater reliability for total score and moderate to high reliability for component scores. Deliberate calibration and training of raters was likely a contributing factor to these results. In fact, the BOSATS scale showed higher inter-rater reliability than both GOALS and OSATS GRS scales. The test-retest reliability was also very high, more than satisfying the predefined cutoff of 0.8 that has been suggested to be acceptable.

The feasibility of use for the BOSATS scale was demonstrated by noting a relatively small time commitment of 30 to 40 minutes for assessment of 1 LGBP video recording. It could be argued that 40 minutes is a long time to spend assessing 1 LGBP case; however, this time investment to objectively identify areas of weakness and to guide deliberate practice outside the operating room is expected to shorten the learning curves and result in cost savings in the operating room. Most of the minimally invasive operating rooms in academic hospitals

Table 6. Spearman’s Rank Correlation Coefficients Between Bariatric Objective Structured Assessment of Technical Skill Scores and Objective Structured Assessment of Technical Skills Global Rating Scale Scores for Different Components of the Gastric Bypass Operation

<table>
<thead>
<tr>
<th>Component of gastric bypass procedure</th>
<th>n</th>
<th>Spearman’s rho</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creation of jejunojejunostomy</td>
<td>27*</td>
<td>0.59*</td>
<td>0.001*</td>
</tr>
<tr>
<td>Creation of gastric pouch</td>
<td>49*</td>
<td>0.48*</td>
<td>0.001*</td>
</tr>
<tr>
<td>Creation of linear stapled gastrojejunostomy</td>
<td>25*</td>
<td>0.70*</td>
<td>0.000*</td>
</tr>
<tr>
<td>Creation of circular stapled gastrojejunostomy</td>
<td>10</td>
<td>0.41</td>
<td>0.245</td>
</tr>
<tr>
<td>Creation of hand-sewn gastrojejunostomy</td>
<td>14*</td>
<td>0.96*</td>
<td>0.000*</td>
</tr>
</tbody>
</table>

*Significant correlations. n, number of cases reviewed.
have the necessary equipment to permit video recording of cases, making this form of assessment feasible and easy to use.

Construct validity for the components of the BOSATS scale was demonstrated by comparing the scores of novice and experienced surgeons for different components of the LGBP operation. Construct validity refers to the degree to which a score can be interpreted as representing the intended underlying construct.\(^3\)\(^1\) Comparison of total BOSATS scores, rather than component scores, between novice and experienced surgeons was not possible because novice surgeons often learned this complex laparoscopic operation 1 component at a time.\(^2\) A trainee would likely achieve proficiency in the JJ before being allowed to attempt creation of the gastric pouch. So, a novice surgeon for the gastric pouch component would likely be an experienced surgeon for the JJ component. In regard to the definitions of novice and experienced surgeons, Ericsson\(^3\)^\(^2\) noted that years of experience, academic standing or rank, and specialty board certification have often been equated to “expertise” and suggested that a better approach may be to determine expertise through reproducibly superior performance (improved treatment outcomes, diagnoses, etc). Taking that recommendation into account, this study used case volumes (<10 and >100 cases) and objective scoring metrics (OSATS GRS score < 28 and ≥ 28) for definitions of novice and experienced surgeons. As a result, the number of novice and experienced surgeons in each group varied depending on the definition used. Using the OSATS GRS score, 5 novice surgeons who performed a circular stapled GJ were reclassified into the experienced surgeon group (Tables 5, 6). These surgeons were minimally invasive surgery fellows who had performed fewer than 10 LGBP operations “skin-to-skin;” however, they would have likely performed other advanced laparoscopic procedures and developed the necessary skills to move their overall technical proficiency closer to the level of an experienced surgeon. Using case-volume criteria, significant differences between experienced and novice groups were seen for 4 of the 5 components analyzed (Table 5). Creation of the gastric pouch is also often the last component of the operation that is taught to a trainee. So, a surgeon would often be proficient in other advanced laparoscopic techniques by the time he or she begins to create the pouch. This statement is supported by the finding of a greater number of surgeons in the experienced group vs the novice group when surgeons are reclassified using OSATS GRS (Tables 5, 6). Consequently, differences between novice and experienced surgeons for the gastric pouch component were small, and a larger sample size would be required to achieve statistical significance.

Concurrent validity is another measure of validity for the BOSATS scale. It is defined as “an evaluation in which the relationship between the test scores and the scores on another instrument purporting to measure the same construct are related.”\(^3\)^\(^2\) The BOSATS scores were correlated with OSATS GRS scores for various components of the operation. Moderate to high correlations were noted for creation of JJ, gastric pouch, and linear

Figure 3. Spearman’s rank correlation between BOSATS and OSATS GRS for creation of hand-sewn GJ anastomosis. BOSATS, Bariatric Objective Structured Assessment of Technical Skill; OSATS GRS, Objective Structured Assessment of Technical Skills Global Rating Scale.
and hand-sewn GJ, suggesting the both scales are measuring components of the same construct — operative performance. No significant correlation was noted for the circular GJ component. This lack of correlation was a result of the relative homogeneity of surgeons who performed circular stapled anastomoses; all surgeons scored higher than 28 on the OSATS GRS scale.

This study had a number of limitations. Missing data points for the Delphi questionnaire were generated when panel members did not rate all the items; however, regardless of how the missing data points were analyzed (replaced by mean, mode, or neutral “3”), Cronbach’s alpha remained above the predetermined consensus cut-off of 0.80. It was also not possible to demonstrate construct validity for all components of the BOSATS scale (ie, “patient positioning,” “access of peritoneal cavity with optical viewing trocar without prior abdominal insufflations,” “triple stapling technique for creation of jejunojejunostomy,” etc), as some aspects of LGBP procedure were not recorded with the intra-abdominal camera view (eg, “patient positioning”), and surgeons who participated in this study did not use all possible approaches to the LGBP. Future projects should focus on demonstration of construct and concurrent validity of the BOSATS scale for those additional components.

Last, it is important to emphasize that the BOSATS scale was designed and validated for use in formative, rather than summative, assessment. At present, it is intended for use in identification of specific areas of weakness, provision of specific targeted feedback, and facilitation of deliberate practice. Additional studies are required to generate specific proficiency cut-off values for use in high-stakes assessment, certification, and recertification.

CONCLUSIONS

The BOSATS scale is a feasible to use, reliable, and valid instrument for objective assessment of operative performance in LGBP. The list of steps generated by the Delphi expert panel created a “training itinerary,” which can be used to guide development of a comprehensive training curriculum for laparoscopic bariatric surgery. Implementation of the BOSATS scale in such a curriculum has the potential to provide trainees with objective structured feedback, facilitate deliberate practice, and shorten learning curves in the operating room.

Author Contributions

Study conception and design: Zevin, Aggarwal, Grantcharov

Acquisition of data: Zevin, Grantcharov

Analysis and interpretation of data: Zevin, Bonrath, Aggarwal, Dedy, Ahmed, Grantcharov

Drafting of manuscript: Zevin

Critical revision: Zevin, Bonrath, Aggarwal, Dedy, Ahmed, Grantcharov

REFERENCES