Assessment of surgery residents’ operative skills in the operating theater using a modified Objective Structured Assessment of Technical Skills (OSATS): A prospective multicenter study

Cornelis J. Hopmans, MD,a Pieter T. den Hoed, PhD,b Lijckle van der Laan, PhD,c Erwin van der Harst, PhD,d Maarten van der Elst, PhD,e Guido H. H. Mannaerts, PhD,f Imro Dawson, PhD,g Reiniër Timman, PhD,h Bas P. L. Wijnhoven, PhD,h and Jan N. M. IJzermans, PhD,a Rotterdam, Breda, Delft, and Capelle aan den IJssel, The Netherlands

Background. With the implementation of competency-based curricula, Objective Structured Assessment of Technical Skills (OSATS) increasingly is being used for the assessment of operative skills. Although evidence for its usefulness has been demonstrated in experimental study designs, data supporting OSATS application in the operating room are limited. This study evaluates the validity and reliability of the OSATS instrument to assess the operative skills of surgery residents in the operating theater.

Methods. Twenty-four residents were recruited from seven hospitals within a general surgical training region and classified equally into three groups according to postgraduate training year (PGY). Each resident had to perform five different types of operations. Surgical performance was measured using a modified OSATS consisting of three scales: Global Rating Scale, Overall Performance Scale, and Alphabetic Summary Scale. Validity and reliability metrics included construct validity (Kruskal-Wallis test) and internal consistency reliability (Cronbach’s α coefficient). Spearman’s correlation coefficients were calculated to determine correlations between the different scales.

Results. Eighteen residents (PGY 1–2 [n = 7]; PGY 3–4 [n = 8]; PGY 5–6 [n = 3]) performed 249 operations. Comparisons of the performance scores revealed that evidence for construct validity depended on the difficulty level of the selected procedures. For individual operations, internal consistency reliability of the Global Rating Scale ranged from 0.93 to 0.95. Scores on the different scales correlated strongly (r = 0.62–0.76, P < .001).

Conclusion. Assessment of operative skills in the operating theater using this modified OSATS instrument has the potential to establish learning curves, allowing adequate monitoring of residents’ progress in achieving operative competence. The Alphabetic Summary Scale seems to be of additional value. Use of the Overall Performance Scale should be reconsidered. (Surgery 2014;156:1078-88.)

From the Department of Surgery,a Erasmus MC University Medical Center; Department of Surgery,b Ikazia Hospital, Rotterdam; Department of Surgery, c Amphia Hospital, Breda; Department of Surgery,d Maasstad Hospital, Rotterdam; Department of Surgery,e Reiniër de Graaf Groep, Delft; Department of Surgery,f Sint Franciscus Gasthuis, Rotterdam; and Department of Surgery,g IJsselland Hospital, Capelle aan den IJssel, The Netherlands

Traditionally, surgical residency training has been based on the master-apprenticeship model. In this training model, characterized by strict obedience to and copying of a role model, residents learned surgical skills almost exclusively in the operating room by working long hours and participating in as many operations as possible. This way of training, however, has been criticized for the subjective assessment of residents’ skills, often based on surrogate markers of operative competence, such as accumulated case
numbers and time taken for a procedure, or biased by personal preferences of supervising surgeons. In addition, the implementation of duty hour regulations has substantially reduced the time that residents are present in the hospital, leaving less time to be spent in the operating theater.

As a result, training opportunities are limited, and operative competence based on excessive exposure to a large variety of operations can no longer be assumed. These developments, in combination with an increased emphasis on debates concerning patient safety and a greater demand for accountability and transparency in medicine, from both a professional and public point of view, have initiated a shift toward a more objective and formalized approach of postgraduate medical training. Within the past decade, many Western countries have restructured the format of their surgical training programs into competency-based curricula. In these curricula, there is a strong need for educational tools that are capable of providing objective performance assessments to evaluate predefined competency goals.

Although operative skills represent only a part of the qualities needed to become a well-trained surgeon, the development of these skills play a pivotal role in any surgical residency program. Within the past decades, various assessment methods have been developed for the evaluation of operative skills, including observational instruments, motion-analysis devices, virtual reality simulations, and the application of computer games. Currently, Objective Structured Assessment of Technical Skills (OSATS) is one of the most applied and accepted observational instruments and is considered as the standard for the assessment of surgical skills.

Originally developed at the University of Toronto, OSATS demonstrated high reliability and validity metrics in a skills laboratory setting on bench model simulations, as well as in animal models and cadaveric training models, suggesting that this instrument could measure operative skills effectively. Subsequently, Datta et al revealed that OSATS might be of value for clinical practice by demonstrating that assessment of technical skills in an inanimate simulation model could be translated to actual surgical performance in the operating room. Since the publication of these results, OSATS has been incorporated in many residency programs, mainly by the specialties of General Surgery and Obstetrics & Gynaecology. As part of a national initiative to reform postgraduate medical training, the Association of Surgeons of the Netherlands also introduced the OSATS instrument into its training program, where it is being used to assess residents’ surgical skills in the operating theater.

To rely on this instrument and to defend its widespread implementation, sufficient evidence comprising metrics for validity and reliability is necessary. However, so far, only a few studies have reported on the use of OSATS in an operating theater setting. In these mainly single-center studies, evidence was either demonstrated by comparing groups with varying experience levels, ie, residents versus surgeons, performing a certain operation, or by evaluating residents in a particular phase of a training program. Therefore, it remains to be demonstrated whether OSATS can discriminate between residents with different training levels. We conducted a multicenter study with the aim to evaluate whether OSATS is a valid and reliable instrument to assess the operative skills of surgery residents in the operating theater.

METHODS

Study design and setting. This prospective observational study was conducted from July 2012 to February 2013 in the operating rooms of seven hospitals within a general surgical training region of the Netherlands, including one university hospital and six district training hospitals. In the Netherlands, the general surgical residency program lasts 6 years and comprises three training phases. The first phase consists of a 2-year training period that is shared with residents from other surgical specialties, including cardiothoracic surgery, orthopaedics, plastic surgery, and urology. Then, a second phase of training is entered that is completely focused on training in general surgery. Finally, residents have to complete 2 years of subspecialty training in one of the following focus areas: gastrointestinal surgery, surgical oncology, trauma surgery, vascular surgery, pediatric surgery, or thoracic surgery. The objective was to assess, using a modified OSATS instrument, each participating resident while he or she performed five different types of operations on six occasions. Assessments could be performed by any board-certified surgeon working at the surgical department in the participating hospitals.

Participants. All program directors (n = 7) of the hospitals within this general surgical training region were asked to list the names of residents eligible to participate in this study, taking into account rotation schedules and the time frame of the study. From these, 24 residents were selected based on postgraduate training year (PGY) and
invited to participate. After receiving informed consent from all invited residents, they were equally classified into three groups, with each group representing one of the different phases of the 6-year training program: group 1 (PGY 1–2), group 2 (PGY 3–4) and group 3 (PGY 5–6).

OSATS. The original OSATS concept consists of a three-part assessment form, including a task-specific checklist, a global rating scale, and a pass/fail judgment. From this original concept, the Association of Surgeons of the Netherlands, bearing responsibility for the format and content of the general surgery residency training program, adopted only the Global Rating Scale, which has shown to be superior in terms of reliability and validity compared with task-specific checklists in a modified form. Modifications included merging of the domains “Knowledge of Instruments” and “Instrument Handling” and adding of the domains “Indication for Surgery” and “Perioperative Management.” Furthermore, the 5-point Likert scale was replaced by a 7-point Likert scale with scores ranging from 4 to 10.

In addition, two new rating scales were added: an Overall Performance Scale and an Alphabetic Summary Scale. On the Overall Performance Scale, both resident and assessor mark a judgment of the performance shown, with awarded scores, ranging from 4 to 10, indicating whether an operation has been performed below, according to, or above expectation. Because this study was aimed at evaluating residents’ operative skills, scores on the Overall Performance Scale assigned by the residents themselves were excluded from data analysis. The Alphabetic Summary Scale provides an assessment of the ability of a resident to perform the operation on that occasion, using a five-level alphabetic scale (level A, assists adequately; level B, competent to perform under strict supervision; level C, competent to perform under limited supervision; Level D, competent to perform without supervision; level E, supervises and educates the procedure). Finally, there is a text box available that can be used by the assessor to motivate the assigned scores on the different OSATS rating scales and to provide feedback on the specific elements of an operation needing improvement. An example of this modified OSATS instrument is shown in Fig 1.

Operations. An overview of the selected operations with prescribed surgical technique is shown in Table 1. These operations were chosen in accordance with the program directors of the participating hospitals for the purpose of this study, because they are common and regularly performed by residents, reflect the broad-based character of general surgical training, and vary in complexity and setting. For example, open and laparoscopic operations were assessed, as well as elective and emergency procedures.

Assessment instructions. Before the study, all residents and attending surgeons of the participating hospitals were briefed on the purpose of the study and received printed instructions on how OSATS assessments should be completed. Residents had to perform the operations as the primary surgeon according to their usual practice and were instructed to register an OSATS assessment of every performed operation selected for this study. Attending surgeons were instructed to act as first assistant, allowing residents to perform the operation within the limits of their competence and to intervene only in the interest of patient safety. Scoring of the modified OSATS instrument was in line with national training directives set by the Association of Surgeons of the Netherlands; this directive prescribes that not all domains of the Global Rating Scale require completion and that residents’ phase of training should be taken into account when scoring the Overall Performance Scale. Finally, it was highlighted that assessments had to be completed directly after the operation, and scores should not be based on previous operating theater experiences with the participating residents.

Statistical analysis. Statistical analyses were performed using IBM SPSS version 20.0 (IBM Corp., Armonk, NY). For all performance scores on the different rating scales of the modified OSATS instrument, mean values and standard deviations were calculated. Because not all domains of the Global Rating Scale require completion during an assessment, scores were based on the average score of the completed domains. To perform statistical analyses, scores on the Alphabetic Summary Scale were interpreted as continuous data and are displayed numerically (score A = 1, score E = 5).

Because our data were not normally distributed, nonparametric tests were used for reliability and validity analyses. Construct validity, defined as the ability of an instrument to differentiate between groups with varying experience levels, was determined by comparing scores on the different rating scales across the three groups of residents using the Kruskal-Wallis test. The Mann-Whitney U test was applied for post-hoc analyses. Cronbach’s α coefficient was determined to establish the internal consistency reliability of the Global Rating Scale. The value of this coefficient reflects to which
Fig 1. Description of the OSATS instrument used by the Association of Surgeons of the Netherlands to assess surgery residents’ operative skills in the operating room. Modified from Martin et al.17
of 540 operations (18 residents were completed with an OSATS assessment, resulting in 18 residents performed 255 operations that resulting from subspecialty training. The remain-

of the increased complexity of rotation schedules indicated they were not able to participate because residents of group 3 (PGY 5–6) withdrew and (PGY 1–2) left the training program, whereas five before the study started. One resident of group 1

RESULTS

relations between the different rating scales. Correlation coefficients were calculated to determine cor-

extent different domains contribute to the overall score on this particular scale. Spearman’s correla-
tions between the different rating scales.

Table I. Overview of the selected operations with prescribed surgical technique

<table>
<thead>
<tr>
<th>Indication for surgery</th>
<th>Surgical technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symptomatic cholecystolithiasis</td>
<td>Laparoscopic cholecystectomy (four-port, two-handed technique)</td>
</tr>
<tr>
<td>Inguinal hernia</td>
<td>Lichtenstein repair</td>
</tr>
<tr>
<td>Inguinal hernia</td>
<td>Totally extraperitoneal procedure</td>
</tr>
<tr>
<td>Proximal femoral fracture</td>
<td>Internal fixation by cannulated hip screws, sliding hip screw, or intramedullary nailing</td>
</tr>
<tr>
<td>Breast tumor</td>
<td>Lumpectomy or mastectomy (eventually followed by sentinel node biopsy or axillary lymph node dissection)</td>
</tr>
</tbody>
</table>

RESULTS

Of the 24 recruited residents, six dropped out before the study started. One resident of group 1 (PGY 1–2) left the training program, whereas five residents of group 3 (PGY 5–6) withdrew and indicated they were not able to participate because of the increased complexity of rotation schedules resulting from subspecialty training. The remaining 18 residents performed 255 operations that were completed with an OSATS assessment, revealing a discrepancy between the total number of 540 operations (18 residents × 5 operations × 6 occasions) expected to be performed within the time frame of this study and the actual number of procedures carried out. Of these 255 operations, six were excluded because of new findings per-operatively (cholecystitis, n = 2) or due to applied surgical technique (hip hemi-arthroplasty, n = 2; axillary lymph node dissection without breast surgery, n = 2). The remaining 249 operations included 80 laparoscopic cholecystectomies, 69 inguinal hernia repairs of which 45 were performed by a Lichtenstein repair and 24 by a totally extraperitoneal procedure (TEP), 41 osteosyntheses of proximal femoral fractures, and 59 breast tumor operations (in some cases followed by sentinel lymph node biopsy or axillary lymph node dissection). All operations were evaluated by 41 attending surgeons who completed a mean number of six (median 4, range 1–21) OSATS assessments per surgeon. The mean number of operations performed per resident was 14 (median 13, range 2–27), with each resident having a mean number of five (median 4, range 2–8) different attending surgeons who completed the OSATS assessments.

Construct validity. For each of the operations, performance scores on the different rating scales of the modified OSATS instrument are presented in Table II. Comparisons of these scores are displayed in Fig 2, A–E. For the laparoscopic cholecystectomy (Fig 2, A), scores on all three rating scales improved across the three groups of residents (P < .005), except for scores on the Overall Performance Scale between group 2 and group 3 (P = .057). Scores for the Lichtenstein repair (Fig 2, B) demonstrated differences on all three rating scales between group 1 and group 2 (P < .001). In contrast, no differences were found between group 2 and group 3. For the TEP (Fig 2, C), differences were present in the scores on the Overall Performance Scale between group 1 and group 2 (P = .036), and in the scores on the Global Rating Scale (P = .003) and the Alphabetic Summary Scale (P = .001) between group 2 and group 3. Analysis of the scores for osteosyntheses of proximal femoral fractures (Fig 2, D) revealed only differences on the Alphabetic Summary Scale between group 1 and group 2 (P = .011). No differences were found in the scores on any of the other rating scales across the groups of resi-

Internal consistency reliability. The internal consistency reliability of the Global Rating Scale for each of the operations was excellent, with Cronbach’s α coefficients ranging from 0.95 to 0.95 (Table II), indicating that scores on all do-

Correlations. Scores on the Global Rating Scale correlated strongly with scores on the Overall Performance Scale (r = 0.76, P < .001) and scores on the Alphabetic Summary Scale (r = 0.75, P < .001). A weaker correlation was observed be-

Table III, correlation coefficients between scores for the individual domains of the Global Rating Scale and the scores on both the Overall Performance Scale and Alphabetic Summary Scale are shown. All correlations coefficients were statistically significant, with all P < .001. The domain “Flow of Operation” (r = 0.78) demonstrated the strongest correlation with scores on
the Overall Performance Scale, whereas the domains “Indication for Surgery” \((r = 0.45)\) and “Perioperative Management” \((r = 0.53)\) showed the weakest correlations. The domains “Time and Motion,” “Knowledge and Handling of Instruments,” and “Use of Assistants” demonstrated the strongest correlations with scores on the Alphabetic Summary Scale. Again, the weakest correlations were found for the domains “Indication for Surgery” \((r = 0.34)\) and “Perioperative Management” \((r = 0.39)\).

**DISCUSSION**

The aim of this study was to examine whether OSATS is a valid and reliable instrument to assess the operative skills of surgery residents in the operating theater. Compared with previous studies reporting on the evaluation of the OSATS instrument in an operating theater environment, we chose a study design in more resemblance with daily clinical practice. The residents participating in this study were selected from all postgraduate training years within a regionally organized general surgery training program and classified into three groups, with each group representing a particular phase of the training program (group 1 PGY 1–2; group 2 PGY 3–4; group 3 PGY 5–6). In this way, classification based on a vast difference in operative experience between the studied groups such as in previous studies was avoided. In addition, it also allowed us to study residents of various training levels compared with other previous studies examining the use of the OSATS instrument in the operating room in which merely residents from certain postgraduate training years were enrolled. Furthermore, a variety of operations were evaluated in comparison with other studies in which only one specific procedure was assessed, such as laparoscopic cholecystectomy or carotid endarterectomy. This approach enabled us to evaluate the validity and reliability of the OSATS instrument to assess operations with varying difficulty levels that are performed commonly by surgery residents during general surgical training in the Netherlands.

**Table II.** Performance scores on the different rating scales of the modified OSATS instrument for the three groups of residents

<table>
<thead>
<tr>
<th>Rating scale</th>
<th>Group 1 (PGY 1–2)</th>
<th>Group 2 (PGY 3–4)</th>
<th>Group 3 (PGY 5–6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 7*</td>
<td>n = 8*</td>
<td>n = 3*</td>
</tr>
<tr>
<td>Laparoscopic cholecystectomy</td>
<td>29</td>
<td>35</td>
<td>16</td>
</tr>
<tr>
<td>GRS ((\alpha = .93))</td>
<td>7.5</td>
<td>0.77</td>
<td>8.4</td>
</tr>
<tr>
<td>OPS</td>
<td>7.7</td>
<td>0.91</td>
<td>8.3</td>
</tr>
<tr>
<td>ASS</td>
<td>2.7</td>
<td>0.55</td>
<td>3.9</td>
</tr>
<tr>
<td>Lichtenstein repair</td>
<td>23</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>GRS ((\alpha = .94))</td>
<td>7.8</td>
<td>0.54</td>
<td>8.7</td>
</tr>
<tr>
<td>OPS</td>
<td>7.8</td>
<td>0.66</td>
<td>8.9</td>
</tr>
<tr>
<td>ASS</td>
<td>2.5</td>
<td>0.51</td>
<td>4.6</td>
</tr>
<tr>
<td>Totally extraperitoneal procedure</td>
<td>10</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>GRS ((\alpha = .93))</td>
<td>7.3</td>
<td>0.35</td>
<td>7.7</td>
</tr>
<tr>
<td>OPS</td>
<td>7.0</td>
<td>0.00</td>
<td>7.7</td>
</tr>
<tr>
<td>ASS</td>
<td>2.6</td>
<td>0.53</td>
<td>2.7</td>
</tr>
<tr>
<td>Osteosynthesis of a proximal femoral fracture</td>
<td>18</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>GRS ((\alpha = .94))</td>
<td>7.8</td>
<td>0.63</td>
<td>8.3</td>
</tr>
<tr>
<td>OPS</td>
<td>8.1</td>
<td>0.88</td>
<td>8.4</td>
</tr>
<tr>
<td>ASS</td>
<td>2.9</td>
<td>0.66</td>
<td>4.0</td>
</tr>
<tr>
<td>Breast tumor surgery</td>
<td>19</td>
<td>28</td>
<td>12</td>
</tr>
<tr>
<td>GRS ((\alpha = .95))</td>
<td>7.2</td>
<td>0.73</td>
<td>8.1</td>
</tr>
<tr>
<td>OPS</td>
<td>7.1</td>
<td>0.64</td>
<td>8.1</td>
</tr>
<tr>
<td>ASS</td>
<td>2.2</td>
<td>0.44</td>
<td>3.2</td>
</tr>
</tbody>
</table>

*Number of residents per group.
|Number of operations performed per group.

\(\alpha\) Internal consistency reliability of the GRS.

ASS, Alphabetic Summary Scale, with scores ranging from 1 [assists adequately] to 5 [supervises and educates]; GRS, Global Rating Scale, with scores ranging from 4 to 10; OPS, Overall Performance Scale, with scores ranging from 4 to 10; PGY, postgraduate training year; SD, standard deviation.
and requires advanced laparoscopic skills which require optimal video-eye-hand coordination, a skill residents master gradually during their training. For the Lichtenstein repair (Fig 2, B), which is one of the first and relatively simple operations residents are exposed to, differences were, as expected, only found between group 1 and group 2. In contrast, for the TEP (Fig 2, C) differences were mainly observed between group 2 and group 3, demonstrating that this operation is complex and requires advanced laparoscopic skills which are being mastered in a later phase of the training program. When performance scores for the osteosyntheses of proximal femoral fractures (Fig 2, D) were compared, almost no differences were found across the different groups of residents. A potential explanation for this lack of construct validity could be the simplicity of this operation. Comparisons of the scores for the breast tumor operations (Fig 2, E) showed differences on almost all rating scales between the three groups of residents. This evidence for construct validity, however, might be biased by the varying difficulty level of this operation, because lumpectomy or mastectomy occasionally was followed by sentinel node biopsy or axillary lymph node dissection. These findings indicate generally that this modified OSATS instrument has the potential to establish learning curves, allowing adequate monitoring of residents’ progress in achieving operative competence.

A more precise examination of the different rating scales of this modified OSATS instrument revealed several interesting findings. First, the Overall Performance Scale demonstrated the least differences, indicating that this rating scale had the weakest discriminative ability. Because scores on this rating scale display whether a resident has performed an operation according to the expectation of the assessor who completes the assessment, we conclude that scores on this scale do not provide an objective indication for operative competence but rather reflect a subjective perception of surgical performance that might be tainted by the personal relationship between assessor and resident. Therefore, use of the Overall Performance Scale should be reconsidered, especially because this rating scale is not part of the original OSATS concept.

In contrast, the Alphabetic Summary Scale revealed the most differences between the groups of residents, indicating that this scale had the greatest discriminative power. This important finding demonstrates the additional value of the Alphabetic Summary Scale and supports its application for use in clinical practice. In our opinion, this rating scale also provides the most objective

<table>
<thead>
<tr>
<th>Domains of the GRS</th>
<th>OPS n</th>
<th>Spearman’s rho</th>
<th>ASS n</th>
<th>Spearman’s rho</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indication for Surgery</td>
<td>183</td>
<td>0.45</td>
<td>184</td>
<td>0.34</td>
</tr>
<tr>
<td>Respect for Tissue</td>
<td>232</td>
<td>0.69</td>
<td>234</td>
<td>0.61</td>
</tr>
<tr>
<td>Time and Motion</td>
<td>231</td>
<td>0.70</td>
<td>233</td>
<td>0.74</td>
</tr>
<tr>
<td>Knowledge and Handling of Instruments</td>
<td>226</td>
<td>0.68</td>
<td>228</td>
<td>0.74</td>
</tr>
<tr>
<td>Use of Assistants</td>
<td>215</td>
<td>0.64</td>
<td>217</td>
<td>0.74</td>
</tr>
<tr>
<td>Flow of Operation</td>
<td>232</td>
<td>0.78</td>
<td>234</td>
<td>0.69</td>
</tr>
<tr>
<td>Knowledge of Specific Procedure</td>
<td>232</td>
<td>0.66</td>
<td>233</td>
<td>0.61</td>
</tr>
<tr>
<td>Perioperative Management</td>
<td>210</td>
<td>0.53</td>
<td>211</td>
<td>0.39</td>
</tr>
</tbody>
</table>

All correlations were statistically significant with all $P < .001$. 

**Fig 2.** Performance scores on the different rating scales of the modified OSATS instrument for each of the operations: (A) laparoscopic cholecystectomy, (B) Lichtenstein repair, (C) totally extraperitoneal procedure (TEP), (D) osteosynthesis of a proximal femoral fracture, (E) breast tumor surgery. Data are presented as means and standard deviations (exact values are depicted in Table II). Results are based on 18 residents: group 1 (PGY 1–2, $n = 7$); group 2 (PGY 3–4, $n = 8$); and group 3 (PGY 5–6, $n = 3$). Left y-axis: GRS (Global Rating Scale) and OPS (Overall Performance Scale) with scores ranging from 4 to 10; Right y-axis ASS (Alphabetic Summary Scale) with scores ranging from 1 to 5. Comparisons between groups, as an indication for construct validity, were made using the Kruskal-Wallis test. Post hoc analyses were performed with the Mann-Whitney $U$ test.
indication for surgical performance because scores reflect whether a resident is considered competent to perform an operation. Such a clear definition of competence is much more informative than quantifying surgical skills by circling scores on numerical rating scales.

When correlation statistics were calculated to evaluate the relationship between scores for the individual domains of the Global Rating Scale and scores on both the Overall Performance Scale and Alphabetic Summary Scale (Table III), stronger correlations were found for the domains related to aspects of operative skill, such as “Respect for Tissue,” “Time and Motion,” “Use of Assistants,” etc. In contrast, weakest correlations were observed for the domains “Indication for Surgery” and “Perioperative Management.” Interestingly, these two domains were added to the original concept of the Global Rating Scale. Because scores on these domains, although critical components of operative care, do not reflect operative skill in any way, we recommend that both domains should be deleted from the modified version of the Global Rating Scale used in this study and should not be adopted by others.

The reliability of this modified OSATS instrument was determined by calculating the internal consistency reliability of the Global Rating Scale, which was excellent for all five operations included in this study. Other reliability metrics that are determined frequently when an assessment tool is evaluated are inter-test reliability and inter-rater reliability. Inter-test reliability reflects the ability of an instrument to generate similar results when a specific operation is evaluated by the same rater on different time points. Inter-rater reliability refers to the extent of agreement between scores when multiple raters evaluate simultaneously the same operation. 26,27 Ideally, to avoid bias when determining these reliability metrics, consecutive operations should be evaluated by multiple raters who are blinded to the training level of the residents being assessed; however, defining a large group of residents who are able to perform a high number of consecutive operations within a particular time frame cannot be realized in this training region because of stringent duty hour regulations.

Moreover, finding attending surgeons who want to participate in such a study will pose an even greater challenge, especially because they have to originate from other hospitals if they need to act as blinded raters. These requirements illustrate that conducting high-quality studies designed to assess surgical skills in an uncontrolled setting like an operating room is a challenging, but above all, extremely difficult undertaking. The loss of residents in the senior group (PGY 5–6) and the considerable discrepancy between the expected number of 540 operations and the actual number of 255 operations that were performed within the time frame of this study exemplify this difficulty.

A remarkable finding in this study were the high values and limited spread of the scores on both the Global Rating Scale and Overall Performance Scale with means ranging from 7.0 to 9.0 (on a 4–10 rating scale). A possible explanation for this finding could be the relative simplicity of the selected operations. However, similar findings also were observed in other studies in which the surgical skills of residents were examined in an operating theater setting, using observational instruments. 28,31,34

Consequently, one might wonder whether high performance scores accurately reflect operative competence. In this context, Williams et al 35 investigated the prognostic value of clinical performance ratings within a general surgery residency program in the United States. They demonstrated that a vast majority of the residents received “Excellent” or “Very Good” scores and only a small minority received “Fair” or “Poor” scores, from which the authors concluded that such ratings have no directly interpretable meaning.

This finding might have important consequences as we move forward into a competency-based assessment model of surgical training. First, high scores make it difficult, to identify residents who are incompetent and function substandardly compared to their peers. Secondly, high scores do not provide residents with any information on the specific elements of an operation that can be improved. In addition, several other issues associated with the evaluation of operative skills in an operating theater environment could cause bias. Because residents and surgeons often closely collaborate in daily practice, it is not unlikely that personal preferences may influence scoring behavior, and thus assessors might be chosen for other reasons than their ability to complete an assessment. Finally, it is almost impossible to perform a standardized assessment for a specific operation because of anatomic variations and varying patient characteristics. These issues open the debate regarding the role OSATS assessments should fulfill in clinical practice, either formative or summative. Formative assessments allow residents and their trainers to monitor progression and can help structuring learning processes by providing constructive feedback.
In contrast, summative assessments have the potential to be used for credentialing or certification purposes. We believe, however, that such important decisions should be based on instruments that provide a valid and reliable indication for operative competence. In a systematic review, addressing the validity and reliability of various methods for surgical skills assessment it was recommended that the OSATS instrument should not be used for summative evaluation purposes due to a lack of high-level evidence. This is an important message that should be taken into account by responsible professional associations and other stakeholders involved in postgraduate medical training who also have adopted the OSATS instrument for surgical skills assessment.

This is the first multicenter study that prospectively evaluated whether OSATS is a valid and reliable instrument to assess the surgical skills of residents enrolled in various phases of a general surgery residency program, while performing a variety of operations in an operating theater setting. Despite that a modified version of the original OSATS concept was used that limited comparison of our findings to those in other studies, we demonstrated that performance scores on the different rating scales improved along with the training level of the residents, which constitutes evidence for construct validity. Evidence for reliability was shown by the high internal consistency reliability coefficients for scores on the Global Rating Scale. Given its great discriminative ability, the Alphabetic Summary Scale seems to be of additional value for clinical practice, while use of the Overall Performance Scale should be reconsidered. Nevertheless, it must be pointed out that, despite observational instruments, like OSATS, may currently be the most accepted and easily available assessment method to evaluate residents’ surgical skills in the operating room, performance scores on these instruments are inherently associated with subjectivity and do not necessarily provide an accurate indication for operative competence.

The authors thank all surgery residents and attending surgeons within general surgeon training region IV in the Netherlands for their participation and cooperation in this study.

REFERENCES