Canadian Medical Education Journal
Revue canadienne de l'éducation médicale

Assessment of laparoscopic skills: Comparing the reliability of global rating and entrustability tools
Évaluation des compétences en laparoscopie : comparaison de la fiabilité des outils d'évaluation globale et des outils d'évaluation de la confiance

Kameela Miriam Alibhai, Amanda Fowler, Nada Gawad, Timothy J Wood et Isabelle Raîche

Volume 13, numéro 6, 2022

URI : https://id.erudit.org/iderudit/1094270ar
DOI : https://doi.org/10.36834/cmej.72369

Aller au sommaire du numéro

Éditeur(s)

Canadian Medical Education Journal

ISSN

1923-1202 (numérique)

Découvrir la revue

Citer cet article

Alibhai, K., Fowler, A., Gawad, N., Wood, T. & Raîche, I. (2022). Assessment of laparoscopic skills: Comparing the reliability of global rating and entrustability tools. *Canadian Medical Education Journal / Revue canadienne de l'éducation médicale*, *13*(6), 36–45. https://doi.org/10.36834/cmej.72369

Résumé de l'article

Contexte : Les programmes de résidence structurés autour de la compétence par conception (CPC) dépendent de plus en plus d'outils qui fournissent des évaluations fiables, nécessitent une formation minimale des évaluateurs et mesurent la progression dans les étapes de la CPC. Pour évaluer les compétences peropératoires, les échelles d'évaluation globale et de confiance sont couramment utilisées mais peuvent nécessiter une formation approfondie. Le Continuum des compétences (CC) est un cadre de la CPC qui peut être utilisé comme outil d'évaluation des compétences laparoscopiques. L'étude visait à comparer le CC à deux autres outils d'évaluation : l'évaluation globale opératoire des compétences laparoscopiques (GOALS) et l'échelle de Zwisch.

Méthodes : Quatre chirurgiens experts ont évalué trente vidéos de cholécystectomie laparoscopique. Deux évaluateurs ont utilisé l'échelle GOALS tandis que les deux autres ont utilisé l'échelle Zwisch et le CC. Chacun d'eux avait reçu une formation spécifique à l'échelle utilisée. Des statistiques descriptives, la fiabilité inter-évaluateurs (FIÉ) et des corrélations de Pearson ont été calculées pour chaque échelle.

Résultats : Des corrélations positives significatives ont été trouvées entre les échelles GOALS et Zwisch ($r=0.75$, $p<0.001$), CC et GOALS ($r=0.79$, $p<0.001$), et CC et Zwisch ($r=0.90$, $p<0.001$). Le CC avait une fiabilité inter-évaluateurs de 0,74 tandis que les échelles GOALS et Zwisch avaient des fiabilités inter-évaluateurs de 0,44 et 0,43, respectivement. Par rapport aux échelles GOALS et Zwisch, le CC avait la fiabilité inter-évaluateurs la plus élevée et ne nécessitait qu'une formation minimale des évaluateurs pour obtenir des scores fiables.

Conclusion : Le CC constituerait un outil fiable pour évaluer les compétences laparoscopiques peropératoires et pour fournir aux stagiaires une rétroaction formatrice pertinente pour les étapes de la CPC. Des recherches supplémentaires devraient être entreprises pour recueillir plus de preuves de validité pour l'utilisation du CC comme outil d'évaluation indépendant.

Canadian Medical Education Journal

Original Research

Assessment of laparoscopic skills: comparing the reliability of global rating and entrustability tools

Évaluation des compétences en laparoscopie : comparaison de la fiabilité des outils d'évaluation globale et des outils d'évaluation de la confiance

Kameela Miriam Alibhai,[1] Amanda Fowler,[2] Nada Gawad,[1,3] Timothy J Wood,[3] Isabelle Raîche[1,3]

[1]Division of General Surgery, Department of Surgery, Faculty of Medicine, University of Ottawa, Ontario, Canada; [2]Division of General Surgery, Department of Surgery, Faculty of Medicine, Memorial University, Newfoundland and Labrador, Canada; [3]Department of Innovation in Medical Education (DIME), University of Ottawa, Ontario, Canada

Correspondence to: Kameela Alibhai, 451 Smyth Rd #2044, Ottawa, ON K1H 8M5; phone: 1-416-843-6247; email: kalib090@uottawa.ca

Published ahead of issue: Aug 2, 2022; published: Nov 15, 2022. CMEJ 2022, 13(6). Available at https://doi.org/10.36834/cmej.72369

Abstract

Background: Competence by design (CBD) residency programs increasingly depend on tools that provide reliable assessments, require minimal rater training, and measure progression through the CBD milestones. To assess intraoperative skills, global rating scales and entrustability ratings are commonly used but may require extensive training. The Competency Continuum (CC) is a CBD framework that may be used as an assessment tool to assess laparoscopic skills. The study aimed to compare the CC to two other assessment tools: the Global Operative Assessment of Laparoscopic Skills (GOALS) and the Zwisch scale.

Methods: Four expert surgeons rated thirty laparoscopic cholecystectomy videos. Two raters used the GOALS scale while the remaining two raters used both the Zwisch scale and CC. Each rater received scale-specific training. Descriptive statistics, inter-rater reliabilities (IRR), and Pearson's correlations were calculated for each scale.

Results: Significant positive correlations between GOALS and Zwisch ($r = 0.75$, $p < 0.001$), CC and GOALS ($r = 0.79$, $p < 0.001$), and CC and Zwisch ($r = 0.90$, $p < 0.001$) were found. The CC had an inter-rater reliability of 0.74 whereas the GOALS and Zwisch scales had inter-rater reliabilities of 0.44 and 0.43, respectively. Compared to GOALS and Zwisch scales, the CC had the highest inter-rater reliability and required minimal rater training to achieve reliable scores.

Conclusion: The CC may be a reliable tool to assess intraoperative laparoscopic skills and provide trainees with formative feedback relevant to the CBD milestones. Further research should collect further validity evidence for the use of the CC as an independent assessment tool.

Résumé

Contexte : Les programmes de résidence structurés autour de la compétence par conception (CPC) dépendent de plus en plus d'outils qui fournissent des évaluations fiables, nécessitent une formation minimale des évaluateurs et mesurent la progression dans les étapes de la CPC. Pour évaluer les compétences peropératoires, les échelles d'évaluation globale et de confiance sont couramment utilisées mais peuvent nécessiter une formation approfondie. Le Continuum des compétences (CC) est un cadre de la CPC qui peut être utilisé comme outil d'évaluation des compétences laparoscopiques. L'étude visait à comparer le CC à deux autres outils d'évaluation : l'évaluation globale opératoire des compétences laparoscopiques (GOALS) et l'échelle de Zwisch.

Méthodes : Quatre chirurgiens experts ont évalué trente vidéos de cholécystectomie laparoscopique. Deux évaluateurs ont utilisé l'échelle GOALS tandis que les deux autres ont utilisé l'échelle Zwisch et le CC. Chacun d'eux avait reçu une formation spécifique à l'échelle utilisée. Des statistiques descriptives, la fiabilité inter-évaluateurs (FIÉ) et des corrélations de Pearson ont été calculées pour chaque échelle.

Résultats : Des corrélations positives significatives ont été trouvées entre les échelles GOALS et Zwisch ($r=0.75$, $p<0.001$), CC et GOALS ($r=0.79$, $p<0.001$), et CC et Zwisch ($r=0.90$, $p<0.001$). Le CC avait une fiabilité inter-évaluateurs de 0,74 tandis que les échelles GOALS et Zwisch avaient des fiabilités inter-évaluateurs de 0,44 et 0,43, respectivement. Par rapport aux échelles GOALS et Zwisch, le CC avait la fiabilité inter-évaluateurs la plus élevée et ne nécessitait qu'une formation minimale des évaluateurs pour obtenir des scores fiables.

Conclusion : Le CC constituerait un outil fiable pour évaluer les compétences laparoscopiques peropératoires et pour fournir aux stagiaires une rétroaction formatrice pertinente pour les étapes de la CPC. Des recherches supplémentaires devraient être entreprises pour recueillir plus de preuves de validité pour l'utilisation du CC comme outil d'évaluation indépendant.

Introduction

As of July 2020, all Canadian General Surgery residency training programs implemented the Competence By Design (CBD) model,[1] an outcome-based medical education framework developed by the Royal College of Physicians and Surgeons of Canada (RCPSC).[2] In residency training programs, the goals of assessment are two-fold. First, they should predict residents' future performance and their readiness to practice independently. This is traditionally determined through summative assessment, which is administered at infrequent intervals and involves standardized tests and scores.[3] Second, assessments should orient the resident toward gaps in their competency, which is a fundamental component of the CBD model.[4] This is accomplished through formative assessments, which consist of frequent observations in low-pressure environments that provide learners with meaningful feedback leading to improved performance.[5] Considering the importance of formative assessment and the frequency at which it should be administered, residency training programs must select assessment tools that produce comparable scores by raters and that provide residents with the type of feedback (i.e., quantitative vs qualitative) to help close competency gaps.

A variety of tools exist to objectively assess surgical residents' technical skills. Global rating scales (GRS), where raters use a rating scale to score a resident's performance as a whole and in several sub-domains, are commonly used in formative assessment.[6] The Global Operative Assessment of Laparoscopic Skills (GOALS) scale (Appendix A) is one such example that assesses residents on their ability to perform minimally invasive surgical procedures.[7] GOALS has been validated to assess residents' intraoperative skills, both in the operating room and in simulated scenarios.[7–9] Compared to scales that provide an overall assessment (i.e., poor, fair, good etc.), numerical scales, such as GOALS, have been shown to be more useful in formative assessment as they provide a learner with numerical feedback, enabling them to quantitatively assess changes in the identified competency gap.[10]

Entrustability scales are another type of assessment tool that have become increasingly common with the shift to competency-based medical education. These behaviourally anchored, ordinal scales identify aspects of a clinical task that a rater is prepared to delegate to a resident without supervision once they have demonstrated a certain level of competence.[11,12] That is, entrustability scales assess the degree to which a task may be completed by a learner with a range of supervision, from complete to none. The Zwisch scale (Appendix B) is one example of an entrustment scale in which residents are assessed on how much guidance they need to perform the critical steps of a surgical procedure.[13] Previous studies have demonstrated evidence for the construct validity of entrustability scale scores in measuring resident autonomy and intraoperative performance.[14] Learners in CBD programs are evaluated multiple times throughout residency, often using entrustability scales, to determine their level of achieved competency.[14]

The Competence Continuum (CC)[15] is a framework developed for CBD programs by the RCPSC that breaks down specialist education into several integrated stages.[15] Throughout training, residents are assessed to determine which entrustable professional activity (EPA) they can or cannot perform. The number of successfully completed EPAs is then used to determine whether a resident has achieved the competence to move to the next stage of training, which are the five stages outlined on the CC. Although EPAs have been widely adopted, they are limited in that a learner who repeatedly cannot complete the task is not necessarily provided with feedback on what areas were lacking and whether their abilities have improved since the previous assessment.[16] As such, instead of assessing a learner on whether they can or cannot complete an EPA, it may be valuable to determine the CC stage in which their abilities would be categorized. The CC framework was not designed to be an independent rating scale; however, considering the fragmented nature of EPAs,[16] it may be valuable for CBD training programs to use it as an assessment tool to provide learners with formative feedback.

In order to provide a resident progressing through a CBD training program with assessment scores that are effective in highlighting gaps in competency, frequent assessments from multiple raters in a short period of time are required.[17] However, in residency training, there are often time constraints and a limited number of available raters. In addition, due to workloads, it is often not possible to administer assessment tools that require extensive rater training procedures to be used correctly. In this context, it is important to know which scales provide a reliable assessment when used by a group of surgeons undergoing minimal rater training. Although the GOALS scale is commonly used to assess laparoscopic skills, it is not clear to what degree it can be used to assess entrustment. Therefore, the purpose of this study was to compare the GOALS scale to the Zwisch scale and CC framework to

determine the extent to which they produce comparable scores when assessing general surgery residents on a simulated porcine laparoscopic cholecystectomy model.

Methods

Study design

This was an exploratory study to compare the GOALS scale, Zwisch scale and CC framework as tools to assess simulated laparoscopic skills. The study was exploratory in nature as it is the first to our knowledge to pilot the CC as an independent assessment scale to provide formative feedback.[18] Institutional Review Board approval was granted by the Ottawa Health Science Network Research Ethics Board (20160278-01H).

Participants & recruitment

A total of four expert raters were recruited as the participants in this study. To be recruited as a participant, the expert raters had to be a general surgeon at The Ottawa Hospital with 1) at least 5 years in practice, 2) subspeciality training in hepatobiliary or minimally invasive surgery and 3) experience assessing surgical residents. Additionally, all four raters who were selected had a particular interested in surgical education and routinely provide trainees with feedback regarding their intraoperative performance.

Materials

A total of 30 dissections of an ex vivo porcine gallbladder off a liver bed were performed and videotaped with the laparoscopic camera between January 2017 and May 2017. The dissections were performed without guidance by 28 general surgery residents and two staff surgeons. Staff surgeons, with varied number of years in practice, were included in addition to residents to try and capture various levels of competence at the staff level. The videotapes, which ranged between four and forty-one minutes in duration, were then digitized and converted into QuickTime videos through the Macintosh application Final Cut Pro. The videotapes were shortened to start when the laparoscopic camera was inserted into the trainer box and to stop after the gallbladder was free from the liver bed. To ensure participant confidentiality the audio track and any footage of the individual operating or their surrounding environment were removed from the videotape. The raters were permitted to fast-forward or rewind the videos as they deemed appropriate.

The GOALS scale uses a five-point Likert scale to assess five domains: depth perception, bimanual dexterity, efficiency, tissue handling, and autonomy. Descriptive anchors are included at points one, three, and five of the scale. For the purposes of this study, raters using GOALS scored each dissection in four of the five domains. The autonomy domain was omitted to ensure participant confidentiality, which has been done in previous studies.[7,19]

The Zwisch scale is an entrustability scale that scores a resident's operative performance as either: 1) Show and Tell; 2) Smart Help; 3) Dumb Help; or 4) No Help. The "Show and Tell" stage describes a resident who first assists and observes the attending; they demonstrate the least amount of earned autonomy. A resident in the "Smart Help" category alternates between surgeon and first assist roles; they demonstrate an increased ability to perform critical steps of the procedure. The "Dumb Help" stage describes a resident who requires active assistance but can execute almost all the steps of the procedure with minimal guidance. Finally, a resident who falls in the "No Help" category can effectively perform the entire procedure and requires only passive assistance from the attending.[13]

The CC is a competency based medical education framework, which describes four key competency milestones in residency: 1) Transition to Discipline; 2) Foundations of Discipline; 3) Core of Discipline; and 4) Transition to Practice.[15] The "Transition to Discipline" describes a resident who has transitioned from medical school and is adjusting to the new learning environment. The "Foundations of Discipline" describes a resident who has covered the broad-based competencies and can now move onto learning more advanced, discipline-specific tasks. The "Core of Discipline" describes a resident who is performing tasks that are expected of a practicing physician at a more supervised or junior level. Finally, the "Transition to Practice" refers to the last few weeks or months of training, where the resident is expected to integrate all their skills and independently apply them in the clinical setting.

Procedure

The four expert raters underwent scale specific training. Rater 1 and 2 received in-depth training on how to use GOALS prior to completing performance assessments of each dissection using the scale. GOALS was the GRS chosen for this study as it has been used extensively and has been considered the gold standard for laparoscopic skills assessment in the general surgery program at our institution. Although the raters were familiar with the scale, the training was offered to ensure it was applied in a manner consistent with the literature, which recently suggests that more extensive rater training is required to properly employ GRSs.[19,20] The GOALS rater training and scoring process was completed in five parts (see Figure 1).

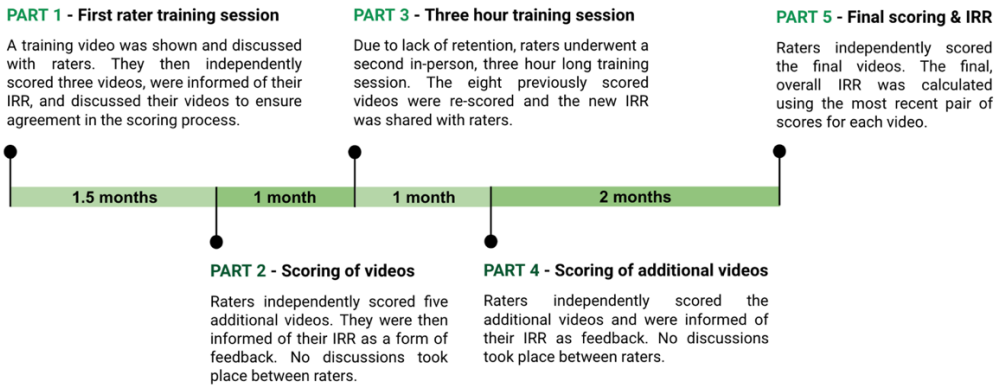

Figure 1. Timeline of the GOALS rater training and scoring process

In Part 1, Raters 1 and 2 underwent an hour and a half long training session where they were shown the rating scale (see Appendix A) and a 12-minute video that highlighted sample dissections and their awarded scores. The raters discussed, amongst themselves and with a member of the research team present who had in-depth knowledge of the GOALS scale, the sample videos as well as the reasons for the assigned scores. The raters then independently scored three videos. The raters were blinded to the identity and level of training of the participant performing the task. Raters 1 and 2 were subsequently informed of the inter-rater reliability (IRR) and discussed their ratings to ensure agreement. This reliability was established using an intraclass correlation (absolute agreement, 2-way mixed-effects model). Part 2 took place one and a half months after; during this session, an additional five videos were scored independently by each rater and raters were again informed of the IRR for all eight videos scored to date for feedback purposes. Due to lack of retention, the raters met in person on month later for a second three-hour long training session with the purpose of achieving standardization (Part 3). During this third part, the eight dissections that had been scored previously were discussed and re-scored by the raters together. The remaining dissections were independently scored either one month or two months after the second training session. In Part 4 and 5, the raters were informed of their new IRR as a form of feedback and did not have any further discussions regarding scoring. The scores from Part 3, 4 and 5 were used for statistical analysis.

Rater 3 and 4 scored all 30 dissections using two tools: the Zwisch scale and the CC. To ensure proper use of the two assessment tools, Raters 3 and 4 received two scale-specific training sessions. The raters were provided with the Zwisch scale (see Appendix B), its anchors and an explanation of the goals of the assessment. Three sample videos were reviewed and discussed. Raters 3 and 4 also received training on how to use the CC, where they were provided a copy of the scale[15] and reviewed the same three sample videos with a member of the research team. Following the training, the two pairs of raters independently scored each video in a random order and were blinded to the identity and level of training of the participant performing the task. Raters 3 and 4 used the Zwisch and CC in the order of their choosing. We chose to limit the formal training for both scales because entrustability tools have been shown to generally require only a small amount of training to produce a reliable result.[21]

Statistical analysis

For each rater, a GOALS score was calculated based on the sum of raters' ratings for each domain (depth perception, bimanual dexterity, efficiency, and tissue handling). A total GOALS score was calculated by averaging the ratings from the two raters. For the Zwisch and CC scales, total scores were based on the mean of the two raters' scores. Descriptive statistics for each scale and the IRR were determined using an ICC (single-rating, consistency, 2-way mixed-effects model). For the subset of videos that were initially scored and then re-scored in Part 3, the ratings from the latter scoring session were used and grouped with the ratings from Part 4 and 5 to calculate the final IRR for GOALS. An ICC of <0.5 was considered poor reliability, 0.5 to 0.75 was moderate, 0.75 to 0.90 was good, and >0.90 was excellent.[22] Pearson's correlation was used to determine the relationship between total scores of the three scales.

Results

A total of 30 laparoscopic cholecystectomies were recorded and de-identified. Table 1 displays the descriptive statistics for each scale. The mean GOALS scores for each rater were similar (*M* = 10.40 and 10.42) and ranged from 6.00 and 16.00 out of a total possible 20.00. The IRR was 0.44 (poor). The mean Zwisch (*M* = 2.63 and 2.50) and CC scores (*M* = 2.70 and 2.80) for each rater were also similar. For Raters 3 and 4, the inter-rater reliability for Zwisch and the CC was 0.43 (poor) and 0 .74 (moderate), respectively.

Table 1. Descriptive statistics for the GOALS, CC and Zwisch scales.

		Mean	± SD	min	max
Goals	Rater 1	10.40	2.36	6.00	16.00
	Rater 2	10.42	2.72	6.00	15.00
	Total	10.41	2.16	6.00	14.00
CC	Rater 3	2.70	1.12	1.00	4.00
	Rater 4	2.80	1.00	1.00	4.00
	Total	2.75	0.99	1.00	4.00
ZWISCH	Rater 3	2.63	1.03	1.00	4.00
	Rater 4	2.50	1.04	1.00	4.00
	Total	2.57	0.88	1.00	4.00

Table 2 demonstrates the frequency of each Zwisch scale score for each rater. Both raters scored the majority of the dissections as either "Active Help" or "Dumb Help."

Table 2. Frequency of Zwisch Scale ratings by rater

Zwisch Rating	**Rater #**	
	Rater 3 Frequency (%)	Rater 4 Frequency (%)
1 – Show and Tell	4 (13.30%)	7 (23.30%)
2 – Active Help	11 (36.70%)	6 (20.00%)
3 – Dumb Help	7 (23.30%)	12 (40.00%)
4 – No Help	8 (26.70%)	5 (16.70%)
Total	30 (100.00%)	30 (100.00%)

Table 3 demonstrates the frequency of CC ratings for each rater. The majority of the dissections were scored as either "Foundations of Discipline" or "Core of Discipline."

Table 3. Frequency of competence continuum framework ratings by rater

CC Rating	**Rater #**	
	Rater 3 Frequency (%)	Rater 4 Frequency (%)
1 – Transition to discipline	5 (16.70%)	3 (10.00%)
2 – Foundations of discipline	9 (30.00%)	9 (30.00%)
3 – Core of discipline	6 (20.00%)	9 (30.00%)
4 – Transition to practice	10 (33.30%)	9 (30.00%)
Total	30 (100.00%)	30 (100.00%)

In comparing ratings between the measures, the Pearson's correlation between the GOALS total score and Zwisch total score was r = 0.75 ($p < 0.001$), between the CC total score and GOALS total score was r = 0.79 ($p < 0.001$), and between CC total score and Zwisch total score was r = 0.90 ($p < 0.001$).

Discussion

The stages of residency outlined in the CC framework have been used since the transition to CBD education, however, the CC has not been studied as a tool to provide residents with feedback on their laparoscopic skills. This was an exploratory study that found that raters' CC scores produced a higher reliability compared to raters' scores using the GOALS and Zwisch scales. Our findings provide preliminary evidence to support the use of the CC framework as a formative laparoscopic skills assessment tool and also suggests that the CC may have superior reliability when compared to two commonly used tools. Training programs may therefore be interested in using the CC as one of their tools of choice for laparoscopic skills assessment as it provides a different type of feedback that is relevant to the CBD milestones, which has been adopted widely among residency programs in Canada.[23]

With the adoption of the CBD model, residency training programs are tasked with selecting a handful of tools from the many assessment tools that can be used to track a resident's progression through the various levels of competence. Having tools that are reliable ensures that the scores learners receive are consistent when administered across different raters and times. The GOALS scale may be useful as it has four items, which allows for a wider range of potential scores and makes it easier to identify small changes in a resident's abilities. Conversely, the Zwisch scale is a single item entrustability scale that requires raters to decide what can safely be delegated to the resident, as is done in the clinical setting on a day-to-day basis.[13,21] While there are no studies that compare how raters employ GRS compared to entrustability scales, the literature suggests that assessment tools that have a greater number of items or contain construct-aligned anchors and narrative wording show greater reliability.[21,24,25] Interestingly, the GOALS has the greatest number of items, yet had an IRR comparable to that of Zwisch (0.44 vs 0.43), which only contains one item. Between Zwisch and CC, which both contain a single item and utilize four construct-aligned anchors, the latter had a higher IRR (0.43 vs 0.74). This finding suggests that even with a single item, raters at our institution were more likely to interpret and employ the CC anchors in similar ways when compared to GOALS and Zwisch. The GOALS' poor IRR may be partly explained by the lack of retention among

raters following the extensive training sessions. Alternatively, the CC's moderate IRR, as compared to Zwisch, may be explained by the fact that the scale anchors demonstrate a higher degree of construct alignment or that it did not require raters to factor in their level of comfort with a resident's technical performance. For example, raters may agree on a resident's global performance according to the CC; however, we hypothesize that a rater who has a high threshold for entrusting residents with certain tasks may be more likely to score them in the Zwisch "Smart Help" category compared to a rater with a lower threshold who may score the same resident in the Zwisch "Dumb Help" category.

Another factor to consider when selecting formative assessment tools is the amount of training necessary to produce a reliable assessment. Initial studies showed that minimal rater training was sufficient to effectively use the GOALS scale;[9] however, recent studies have found poor inter-rater reliabilities when no or minimal rater training was provided.[7,20,26] Given these findings, more extensive rater training was provided to determine its effects on IRR.[7,20,26] Nevertheless, after multiple in depth training sessions, the raters in our study, who were already familiar with GOALS, only achieved poor reliability (r = 0.44). This may suggest that even with extensive and continuous rater training among raters who have experience using GOALS, moderate or good IRR may be difficult to achieve. The lack of IRR may be inherent to GOALS itself and its use of a Likert scale, where raters may be more likely to be on opposite ends of the numeric pole.[7,27] Similarly, the Zwisch scale had poor reliability (r = 0.43) following minimal rater training, which was purposefully limited given recent studies have indicated entrustability scales require minimal rater training to produce reliable assessments. Given the reliability of Zwisch identified in this study, raters may benefit from additional training, especially if they, like Rater 3 and 4, have previously not used the scale.[7,19,20,28] It is unrealistic to expect that multiple expert surgeons would undergo prolonged training sessions to learn how to employ a rating scale.[19,21] For these reasons, the CC framework may be a valuable feedback tool for training programs, as it employs anchors that appear to be easier for raters at our institution to understand and only requires a brief training session to produce a reliable assessment.[21,24]

To effectively assess residents, training programs should try to identify which types of feedback are most valuable to help continuously encourage trainees to develop their competency skills. The GOALS scale provides unique numerical data that can be plotted over time to quantitatively track progression and compare residents to one another. Unfortunately, numeric scores alone provide little constructive feedback that enable residents to continuously hone their abilities.[29,30] As Rekman et al. highlighted, it may be useful for programs to provide entrustment-based feedback instead, as it provides residents with a better understanding of the tasks they can safely perform alone in the clinical setting.[21] Considering the advantages inherent to each type of feedback, residency programs will likely need to use a combination of both numeric and entrustment-based feedback, and the CC framework may be one such tool that can be adopted to provide the latter.

Future research may consider employing a qualitative study design to explore residents' and program directors' impression of the utility of the CC scale as a formative feedback tool. Specifically, it would be of value to identify the type of formative feedback, whether a numerical score, an entrustment score (i.e., dumb help stage) or a position on the CC (i.e., core of discipline), that is most useful in identifying gaps in competence and improving resident performance. If findings from a qualitative study support the use of the CC as a formative feedback tool, it would provide preliminary validity evidence to support the CC being formally implemented into medical curricula as an independent rating scale.

Limitations

Our study is not without its limitations. Due to time constraints and rater availability, the reliability of each scale was calculated using scores from two raters. Having multiple raters would have made it harder for a single rater with extreme scores to influence the reliability, especially for Likert scales. This limited sample does, however, more realistically reflect the time and resource constraints that residency training programs face.[31] In these situations, where residents can only be assessed in a limited number of situations, it is imperative that the assessment tools that are employed provide a reliable assessment with only a few raters. Our small sample size of 30 dissections also limited our ability to capture the entire spectrum of intraoperative competence; however, the level of training among residents and the years in practice among the staff who performed the dissections were varied to ensure a range of abilities were included.

Second, given the recent findings that indicate minimal rater training is likely to lead to a low IRR for GOALS scores, Raters 1 and 2 received multiple in-depth training sessions in a proactive attempt to achieve good reliability. The Zwisch scale was employed with minimal rater training as recent studies have suggested that entrustability scales can be reliably used as such. The CC has not been used as an independent assessment tool before but given it is an entrustment-based framework, minimal rater training was also provided to be able to compare it more accurately to Zwisch, the formal entrustability tool in this study. Consequently, there was great discrepancy in the amount of training provided to employ GOALS when compared to Zwisch and the CC. This does support the argument that even with extensive and ongoing rater training, GOALS may still yield poor IRR and that not all entrustability scales (i.e., Zwisch) can be reliably used with minimal rater training.[96,19]

Finally, the order in which the Zwisch and CC were used was left to the discretion each rater and was therefore not consistent. If, for example, Rater 3 employed the Zwisch scale first, the wording of the Zwisch anchors and the rater's interpretation of them may have influenced the subsequent use and interpretation of the CC's anchors.

Conclusion

This exploratory study was a first step in demonstrating that the CC framework may be an effective tool to provide formative feedback to learners regarding their laparoscopic skills. Compared to GOALS and Zwisch, the CC was employed using minimal rater training and yielded the highest IRR of all three scales. These findings suggest that the CC should be further studied as a reliable tool that can be administered with minimal training.

Conflicts of Interest: None to declare.
Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Royal College of Physicians and Surgeons of Canada. *CBD start, launch and exam schedule.* http://www.royalcollege.ca/rcsite/documents/cbd-community-touchpoint/cbd-rollout-schedule-e. [Accessed Oct 3, 2020].
2. University of Toronto Department of Medicine. *Competency based medical education.* https://www.deptmedicine.utoronto.ca/competency-based-medical-education . [Accessed Oct 3, 2020].
3. de Montbrun S, Satterthwaite L, Grantcharov TP. Setting pass scores for assessment of technical performance by surgical trainees. *Br J Surg*. 2016;103(3):300-306. https://doi.org/10.1002/bjs.10047
4. Epstein, Ronald M., Cox, Malcolm, Irby DM. Assessment in medical education. *NEJM*. 2007;100(2):387-396. https://doi.org/10.1056/nejmra054784
5. Rudolph JW, Simon R, Raemer DB, Eppich WJ. Debriefing as formative assessment: Closing performance gaps in medical education. *Acad Emerg Med*. 2008;15(11):1010-1016. https://doi.org/10.1111/j.1553-2712.2008.00248.x
6. Middleton RM, Baldwin MJ, Akhtar K, Alvand A, Rees JL. Which Global Rating Scale? *J Bone Jt Surg*. 2016;98(1):75-81. https://doi.org/10.2106/JBJS.O.00434
7. Kramp KH, Van Det MJ, Hoff C, Lamme B, Veeger NJGM, Pierie JPEN. Validity and reliability of Global Operative Assessment of Laparoscopic Skills (GOALS) in novice trainees performing a laparoscopic cholecystectomy. *J Surg Educ*. 2015;72(2):351-358. https://doi.org/10.1016/j.jsurg.2014.08.006
8. Gumbs AA, Hogle NJ, Fowler DL. Evaluation of resident laparoscopic performance using global operative assessment of laparoscopic skills. *J Am Coll Surg*. 2007;204(2):308-313. https://doi.org/10.1016/j.jamcollsurg.2006.11.010
9. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg*. 2005;190(1):107-113. https://doi.org/10.1016/j.amjsurg.2005.04.004
10. Williams RG, Sanfey H, Chen XP, Dunnington GL. A controlled study to determine measurement conditions necessary for a reliable and valid operative performance assessment: a controlled prospective observational study. *Ann Surg*. 2012;256(1):177-187. https://doi.org/10.1097/SLA.0b013e31825b6de4
11. YouTube. *March 27 2015 CBD Webinar EPA Milestones.* https://www.youtube.com/watch?v=CScsSywOaAU. [Accessed Oct 3, 2020].
12. The Royal College of Physicians and Surgeons of Canada. *EPAs and CanMEDS milestones.* http://www.royalcollege.ca/rcsite/cbd/implementation/cbd-milestones-epas-e. [Accessed Oct 3, 2020].
13. Darosa DA, Zwischenberger JB, Meyerson SL, et al. A theory-based model for teaching and assessing residents in the operating room. *J Surg Educ*. 2013;70(1):24-30. https://doi.org/10.1016/j.jsurg.2012.07.007
14. George BC, Teitelbaum EN, Meyerson SL, et al. Reliability, validity, and feasibility of the zwisch scale for the assessment of intraoperative performance. *J Surg Educ.*. Vol 71. Elsevier Inc.; 2014:e90-e96. https://doi.org/10.1016/j.jsurg.2014.06.018
15. Royal College of Physicians and Surgeons of Canada. *CBD competence continuum diagram.* 2015;(June):2015. https://www.royalcollege.ca/rcsite/documents/cbd/cbd-competence-continuum-diagram-legal-e.pdf. [Accessed on May 17, 2022].
16. Al-Moteri M. Entrustable professional activities in nursing: a concept analysis. *Int J Nurs Sci*. 2020;7(3):277-284. https://doi.org/10.1016/j.ijnss.2020.06.009
17. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. *Med*

Teach. 2010;32(8):676-682. https://doi.org/10.3109/0142159X.2010.500704
18. Research Guides at University of Southern California. *Types of research designs - organizing your social sciences research paper* https://libguides.usc.edu/writingguide/researchdesigns. [Accessed Mar 11, 2022].
19. Gawad N, Fowler A, Mimeault R, Raiche I. The inter-rater reliability of technical skills assessment and retention of rater training. *J Surg Educ*. 2019;76(4):1088-1093. https://doi.org/10.1016/j.jsurg.2019.01.001
20. Ilgen JS, Ma IWY, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Med Educ*. 2015;49(2):161-173. https://doi.org/10.1111/medu.12621
21. Rekman J, Gofton W, Dudek N, Gofton T, Hamstra SJ. Entrustability scales: outlining their usefulness for competency-based clinical assessment. *Acad Med*. 2016;91(2):186-190. https://doi.org/10.1097/ACM.0000000000001045
22. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-163. https://doi.org/10.1016/J.JCM.2016.02.012
23. Touchie C, Ten Cate O. The promise, perils, problems and progress of competency-based medical education. *Med Educ*. 2016;50(1):93-100. https://doi.org/10.1111/medu.12839
24. Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. *Med Educ*. 2011;45(6):560-569. https://doi.org/10.1111/j.1365-2923.2010.03913.x
25. Weller JM, Castanelli DJ, Chen Y, Jolly B. Making robust assessments of specialist trainees' workplace performance. *Br J Anaesth*. 2017;118(2):207-214. https://doi.org/10.1093/bja/aew412
26. Gawad N, Fowler A, Mimeault R, Raiche I. The inter-rater reliability of technical skills assessment and retention of rater training. *J Surg Educ*. 2019;76(4):1088-1093. https://doi.org/10.1016/j.jsurg.2019.01.001
27. Gallagher AG, Ritter EM, Satava RM. Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. *Surg Endosc Other Interv Tech*. 2003;17(10):1525-1529. https://doi.org/10.1007/S00464-003-0035-4/TABLES/2
28. Bilgic E, Watanabe Y, McKendy K, et al. Reliable assessment of operative performance. *Am J Surg*. 2016;211(2):426-430. https://doi.org/10.1016/j.amjsurg.2015.10.008
29. Tekian A, Watling CJ, Roberts TE, Steinert Y, Norcini J. Qualitative and quantitative feedback in the context of competency-based education. *Med Teach*. 2017;39(12):1245-1249. https://doi.org/10.1080/0142159X.2017.1372564
30. Silber CG, Nasca TJ, Paskin DL, Eiger G, Robeson M, Veloski JJ. Do global rating forms enable program directors to assess the ACGME competencies? *Acad Med*. 2004;79(6):549-556. https://doi.org/10.1097/00001888-200406000-00010
31. Anderson PAM. Giving feedback on clinical skills: are we starving our young? https://doi.org/10.4300/JGME-D-11-000295.1

Appendices

Appendix A. Global rating scale component of the intraoperative assessment tool (GOALS) adapted from Vassiliou 2005[9]

Depth Perception				
1	2	3	4	5
Constantly overshoots target, wide swings, slow to correct.		Some overshooting or missing of target, but quick to correct.		Accurately directs instruments in the correct plane to target.
Bimanual Dexterity				
1	2	3	4	5
Uses only one hand, ignores nondominant hand, poor coordination between hands.		Uses both hands but does not optimize interaction between hands.		Expertly uses both hands in a complimentary manner to provide optimal exposure.
Efficiency				
1	2	3	4	5
Uncertain, inefficient efforts; many tentative movements; constantly changing focus or persisting without progress.		Slow, but planned movements are reasonably organized.		Confident, efficient, and safe conduct, maintains focus on task until it is better performed by way of an alternative approach.
Tissue Handling				
1	2	3	4	5
Rough movements, tears tissue, injures adjacent structures, poor grasper control, grasper frequently slips.		Handles tissues reasonably well, minor trauma to adjacent tissue (i.e., occasional unnecessary bleeding or slipping of the grasper).		Handles tissues well, applies appropriate traction, negligible injury to adjacent structures.
Autonomy				
1	2	3	4	5
Unable to complete entire task, even with verbal guidance.		Able to complete task safely with moderate guidance.		Able to complete task independently without prompting.

Appendix B. The Zwisch Model for teaching and assessment in the operating room adapted from DaRosa 2013[4]