Original Study

ANNALS OF SURGERY OPEN

OPEN

Discrimination, Reliability, Sensitivity, and Specificity of Robotic Surgical Proficiency Assessment With Global Evaluative Assessment of Robotic Skills and Binary Scoring Metrics

Results From a Randomized Controlled Trial

Ruben De Groote, MD,*†‡ Stefano Puliatti, MD,*§ Marco Amato, MD,*§ Elio Mazzone, MD,‖¶ Alessandro Larcher, MD,‖¶ Rui Farinha, MD,* Artur Paludo, MD,*# Liesbeth Desender, MD, PhD,** Nicolas Hubert, MD,†† Ben Van Cleynenbreugel, MD, PhD,‡ Brendan P. Bunting, PhD,‡‡ Alexandre Mottrie, MD, PhD,*† and Anthony G. Gallagher, PhD, DSc, MAE*‡§§

On behalf of the Junior ERUS/ YAU working group on robot-assisted surgery of the European Association of Urology and the ERUS Education Working Group. Collaborators:
Giuseppe Rosiello, MD,*‖¶ Pieter Uvin, MD, PhD,‖‖ Jasper Decoene, MD,¶¶ Tom Tuyten, MD,## Mathieu D'Hondt, MD,*** Charles Chatzopoulos, MD,††† Bart De Troyer, MD,‡‡‡ Filippo Turri, MD,§ Paolo Dell'Oglio, MD,§§§ Nikolaos Liakos, MD,‖‖‖ Carlo Andrea Bravi, MD,*‖¶ Edward Lambert, MD,† Iulia Andras, MD,¶¶¶ Fabrizio Di Maida, MD,### and Wouter Everaerts, MD, PhD****

Objective: To compare binary metrics and Global Evaluative Assessment of Robotic Skills (GEARS) evaluations of training outcome assessments for reliability, sensitivity, and specificity.
Background: GEARS–Likert-scale skills assessment are a widely accepted tool for robotic surgical training outcome evaluations. Proficiency-based progression (PBP) training is another methodology but uses binary performance metrics for evaluations.
Methods: In a prospective, randomized, and blinded study, we compared conventional with PBP training for a robotic suturing, knot-tying anastomosis task. Thirty-six surgical residents from 16 Belgium residency programs were randomized. In the skills laboratory, the PBP group trained until they demonstrated a quantitatively defined proficiency benchmark. The conventional group were yoked to the same training time but without the proficiency requirement. The final trial was video recorded and assessed with binary metrics and GEARS by robotic surgeons blinded to individual, group, and residency program. Sensitivity and specificity of the two assessment methods were evaluated with area under the curve (AUC) and receiver operating characteristics (ROC) curves.
Results: The PBP group made 42% fewer objectively assessed performance errors than the conventional group ($P < 0.001$) and scored 15% better on the GEARS assessment ($P = 0.033$). The mean interrater reliability for binary metrics and GEARS was 0.87 and 0.38, respectively. Binary total error metrics AUC was 97% and for GEARS 85%. With a sensitivity threshold of 0.8, false positives rates were 3% and 25% for, respectively, the binary and GEARS assessments.
Conclusions: Binary metrics for scoring a robotic VUA task demonstrated better psychometric properties than the GEARS assessment.

Keywords: Basic skill training, GEARS, proficiency-based progression, robotic surgery, training

*From the *ORSI Academy, Ghent, Belgium; †Department of Urology, OLV, Aalst, Belgium; ‡Department of Development and Regeneration, KU Leuven, Leuven, Belgium; §Department of Urology, University of Modena and Reggio Emilia, Modena, Italy; ‖Division of Oncology/Unit of Urology, URI, IRCCS Ospedale San Raffaele, Milan, Italy; ¶Vita-Salute San Raffaele University, Milan, Italy; #Clinic Hospital of Porto Alegre, Urology, Porto Alegre, Brazil; **Department of Thoracovascular Surgery, University Hospital Ghent, Ghent, Belgium; ††Department of Urology, CHR de la Citadelle, Liège, Belgium; ‡‡School of Psychology, Ulster University, Coleraine, Northern Ireland, United Kingdom; §§School of Medicine, Faculty of Life and Health Sciences, Ulster University, Northern Ireland, United Kingdom; ‖‖Department of Urology, AZ Sint-Jan, Bruges, Belgium; ¶¶Department of Urology, OLV van Lourdes Hospital, Waregem, Belgium; ##Department of Urology, Jessa Hospital, Hasselt, Belgium; ***Department of Surgery, AZ Groeninge, Kortrijk, Belgium; †††Department of Urology, Chirec Hospital, Brussels, Belgium; ‡‡‡Department of Urology, AZ Nikolaas, Sint-Niklaas, Belgium; §§§Department of Urology, Niguarda Hospital, Milan, Italy; ‖‖‖Prostate Center Northwest, Department of Urology, Pediatric Urology and Uro-Oncology, St. Antonius-Hospital, Gronau, Germany; ¶¶¶Department of Urology, Iuliu Hatieganu University of Medicine and Pharmacy, Cluj-Napoca, Romania; ###Department of Urology, University of Florence, Florence, Italy; and ****Department of Cellular and Molecular Medicine, KU Leuven, Leuven, Belgium.*

EAU Robotic Urology Section (ERUS) Scientific Committee sponsorship of the 7 Belgian surgeons should be acknowledged.

The authors declare that they have nothing to disclose.

R.D.G. did conceptualization, data curation, investigation, methodology, validation, writing (original draft), and writing (review and editing). S.P. and M.A. did conceptualization, data curation, investigation, methodology, validation, and writing (review and editing). E.M. did data curation, Investigation, and writing (review and editing). A.L. did investigation an writing (review and editing). R.F. did data curation, investigation, and writing (review and editing). A.P. did data curation, investigation, and writing (review and editing). L.D. did conceptualization, methodology, and writing (review and editing). N.H. did data curation, investigation, and writing (review and editing). B.V.C. did conceptualization, methodology, and writing (review and editing). B.P.B. did data curation, investigation, methodology, and validation. A.M. did conceptualization, investigation, methodology, validation, and writing (review and editing). A.G.G. did conceptualization, data curation, investigation, methodology, validation, writing (original draft), and writing (review and editing). Collaborators (G.R., P.U., J.D., T.T., M.D'H., C.C., B.D.T., F.T., P.D.'O., N.L., C.A.B., E.L., I.A., F.D.M., and W.E.) did data collection and writing (review and editing).

Robotic surgery has been increasingly used over the past two decades because it combines the advantages of minimally invasive surgery with three-dimensional vision, shorter learning curve, increased dexterity and precision, and ergonomics for the surgeon.[1,2] As surgical skill is related to patient outcome,[3] the implementation of validated training curricula in robotic surgery is key to bring trainees to proficiency and to increase patient safety.[4–8] To discriminate surgical quality and to assess robotic surgical technical skills during and at the completion of training, several assessment tools have been developed and validated.[9] The Global Evaluative Assessment of Robotic Skills (GEARS) was the first surgical technical skills assessment tool specifically for robotic surgery.[10] Since its development, GEARS has become the most extensively studied and applied assessment tool for robotic surgery.[11] However, GEARS is a global assessment tool and not procedure specific. Also, it is a quantitative assessment tool based on Likert-type scales and there is evidence that it may be prone to weak interrater reliability.[12] This is an important issue because an assessment tool with peer reviewed and published validation evidence that is demonstrated to be unreliable is by default not valid.[13]

Ethical concerns about training on living patients, an increase in the number of surgical procedures and their complexity, the financial burden of increased operative time during training procedures and restriction on working hours have forced the surgical community to explore new and more effective and efficient ways of training surgical skills. Proficiency-based progression (PBP) is a training method that has demonstrated its value in different surgical specialties.[8,14] A specific level of training outcome, defined by a quantitative score based on the objectively assessed performance of experienced and practicing surgeons (i.e., a benchmark) must be demonstrated to gain the proficiency level.[15] The cornerstone of PBP training are the procedure specific, validated, binary, and quantitative performance metrics, which are used to train (i.e., formative feedback to the trainee) and assess trainees. During the PROVESA trial,[16] GEARS and Binary metrics were used to assess performance of a robotic suturing and knot-tying anastomosis task, that is, a vesico-urethral anastomosis (VUA) on a chicken model between two groups of surgical residents.[17] The first aim of this study is to evaluate the capacity of both assessment methods to discriminate between the objectively assessed performance of the two groups of surgical trainees learning the robotic suturing and VUA knot-tying task, that is, a PBP-trained group and a conventionally trained group. A second aim was to evaluate psychometric properties (i.e., the reliability, sensitivity, and specificity) of the GEARS and binary checklist assessment tools.

MATERIALS AND METHODS

The PROVESA trial is a multicentric, prospective, randomized, blinded controlled trial during, which conventional training (apprenticeship model) is compared with PBP training for robotic suturing and knot-tying VUA skill training (Figure 1).

SDC *Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.annalsofsurgery.com).*

Reprints: Ruben De Groote, MD, FEBU, Department of Urology, Uro-oncology and Robotics, Onze Lieve Vrouw Ziekenhuis Aalst, Moorselbaan 164, 9300 Aalst, Belgium. Email: degroote.ruben@gmail.com.

Annals of Surgery Open (2023) 3:e307

Received: 1 May 2023; Accepted 3 June 2023

Published online 16 August 2023

DOI: 10.1097/AS9.0000000000000307

Participants/Subjects

The study subjects were 36 first- and second-year residents from 16 training sites and 12 residency training programs (ethical committee and the trial was registered at the National Institution of Health [NCT04786834]). All subjects completed informed consents.

Faculty Training

The PROVESA PBP faculty consisted of 7 surgeons, including 3 consultant urologists (R.D.G., S.P., R.F.) and 4 senior clinical fellows uro-oncology and robotics (A.P., G.R., E.M., M.A.). They were supervised by the PROVESA research coordinator, a consultant behavioral scientist (A.G.G.). During assessment training of raters (prior to commencement of the trial) mastery of the metrics (in Appendix 1, http://links.lww.com/AOSO/A226) was demonstrated by repeated interrater reliability >0.8 during assessment of full-length surgical videos.

Based on their demographic information and performance during the baseline assessment a matched 1:1 randomization was performed (using an online randomizer www.random.org). Subjects were matched for age (±2 years), residency year, surgical specialty, and skill at baseline as determined by the objectively assessed performance metrics score (Figure 1A,B). The Traditional Trained Group (TTG) were yoked to the same training time as their counterpart in the PBP Group.

Group PBP Training

Eighteen participants who were randomized to the PBP group were given access to a dedicated PBP e-course on the online Bridge platform 1 week before their training in the skills lab of ORSI Academy (Figure 1A,B). In this e-course, the operative metrics were reviewed during which performance errors and steps were illustrated. Immediate feedback was given to the participants during the course. Before continuing their training in the skills lab, all subjects had to pass a test by reaching the preset proficiency benchmark (94%) on the eLearning module (defined as the mean score on the test by the panel of PBP experts). Training and assessment methods, including the VUA task are described in more detail in Supplementary information (in Appendix 2, http://links.lww.com/AOSO/A226).

In the skills lab, training was given in teams of 3 participants per trainer during a full day by the PROVESA PBP faculty in a standardized way with the operative metrics as guiding instrument. While one subject was training, the other two participants scored and gave metric-base formative feedback to their colleague on task completion. Emphasis was placed on the different steps the participants had to perform to complete the procedure, but even greater emphasis was placed on performance errors they could make in each step. Each group of 3 trainees was supervised by an experienced faculty member who gave ongoing, formative, metric-based performance feedback based on the performance metrics agreed by procedure experts (i.e., deliberate practice training).[18] The estimation of proficiency during training was not the result of the subjective interpretation of the PBP trainer but was based on the objective metrics-based score given by the PBP trainer to the trainee based on their performance. Readiness for final assessment, as judged by the PBP trainer occurred when the trainee repeatably (i.e., twice consecutively) scored equal to or better than the quantitatively defined proficiency benchmark. Pretrial training and evaluation of faculty in the reliable and accurate use and application of the metrics precluded personal and subjective opinion about the trainee's skillset. The trainee either demonstrated the requisite performance characteristics or they did not. Subsequently, the final trial was supervised, and videorecorded by one of the designated PROVESA faculty members.

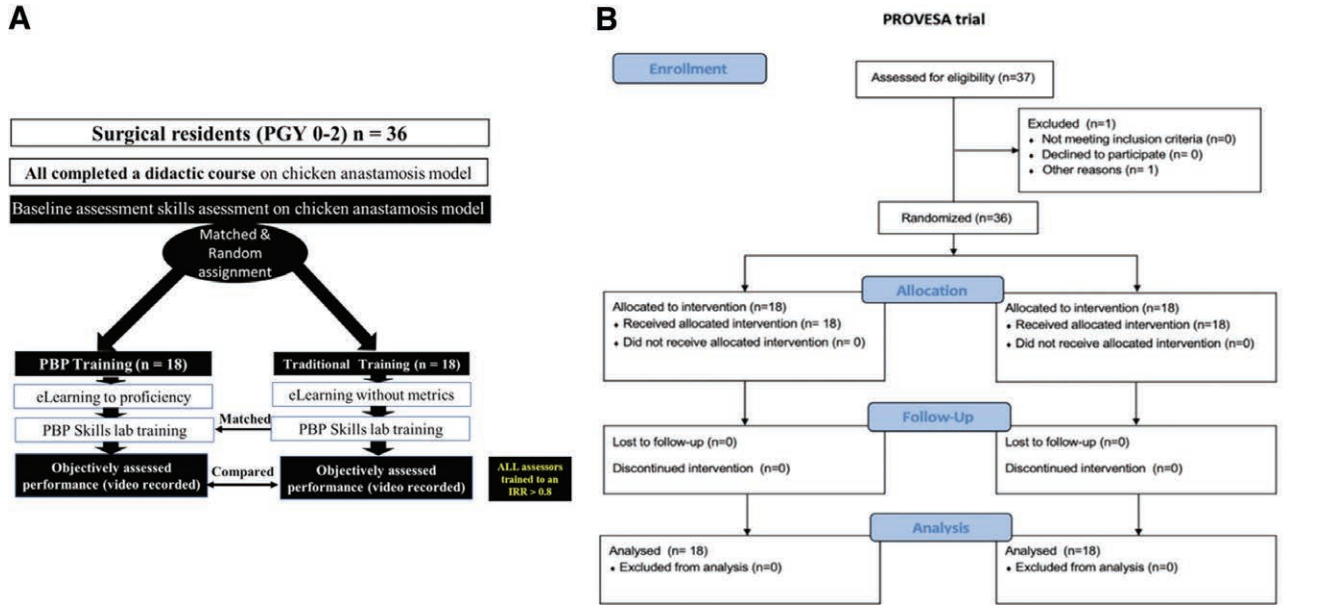

FIGURE 1. A, The design of the multicentre, prospective, randomized, matched, and blinded study and (B) the CONSORT 2010 flow diagram of the PROVESA trial.

Group—Traditional Training

Eighteen participants were randomly assigned to the TTG. As control group, they represent the current standard of care in training. For 1 week, they had continuous access to the exact same online learning platform as the PBP Group and were repeatedly encouraged to study the material. Subsequently, they were invited to ORSI Academy for a full day of training by 7 robotic experts. All experts were selected from different Belgian hospitals and were considered to be experts in robotic surgery and excellent trainers (as evaluated by their peers and residents). All of them had performed >300 robotic procedures independently and were experienced in training residents and/or fellows. These experts were allowed to train as they would do in their own hospital. In the traditional group, the feedback was purely depending on the expertise and training skills of the trainer. The trainee learned by absorbing these tips and tricks and by repeating the same task (i.e., repeated practice).[19] There was also a faculty of 7 different surgeons. There was no systematic consensus between them on how to train the trainees and which tips needed to be given during training.

As matching was done for training time with their counterparts in the PBP group, every participant in the TTG had a preset number of trials before doing the final assessment which was videorecorded. After final assessment, all participants of the TTG completed the same online assessment on the Bridge platform as their counterparts in the PBP group did.

Video Scoring

The 36 full-length study videos, each with only the designated unique identifying number attached were randomly assigned to one of four pairs of reviewers. All video reviewers were blinded to the source of the video (i.e., training group, the trainee identity, hospital and residency training program) being reviewed. Each video was independently reviewed and scored by the 2 assessors.

Final performance was assessed at first using the previously reported,[17] binary operative performance metrics (in Appendix 1, http://links.lww.com/AOSO/A226), which represent a comprehensive overview of the different procedural steps and errors of the specific procedure. The metrics were developed during a procedural characterization by extensive video review. Subsequently, the metrics were presented and discussed at a modified Delphi consensus meeting. Performance metrics consisted of 5 explicitly defined surgical steps (posterior wall, left lateral wall, right lateral wall, anterior wall and knotting) and 20 performance errors (i.e., content validation, generalization, extrapolation, and implication validation level evidence[20–22]). Of these errors, 3 were designated as critical errors because either (1) the error's enactment had the potential to seriously compromise the success of the procedure or (2) the error had the potential to create significant iatrogenic damage to the VUA (if replicated in a real patient). The proficiency benchmark was based on the mean of the objectively assessed performance of the experts during the construct validity study.[17] Proficiency was demonstrated by completing all 5 steps within 25 minutes with 10 or less performance errors and no critical errors. Suture breakage during the task led to immediate failure. The reviewer scored (in a binary fashion) performance units that were or were not observed to have occurred.

Additionally, surgical performance was scored by video review by two faculty using GEARS.[11] The different domains of the GEARS scale were addressed by indication on the Likert scales (1–5) at the discretion of the assigned reviewer.

Statistical Analysis

Statistical analysis was performed with SPSS 26 (Armonk, New York). Differences between the Groups for Binary Checklist Metrics and GEARS assessments were compared with one-factor analysis of variance (ANOVA). The dependent variable for the ANOVA was the GEARS score on final repetition. Differences in the number of trainees in the two groups demonstrating the proficiency benchmark were assessed with a Chi-Square test. Sensitivity (i.e., correctly' able to identify individuals who have demonstrated proficiency) and Specificity (correctly' able to identify individuals who have failed to demonstrate proficiency) of the metrics were assessed with area under the curve (AUC) of receiver operating characteristics (ROC) curves and asymptotic significance tests.

Score Tabulation

For the entire procedure, the total number of steps completed, errors made, and critical errors enacted were averaged for the pair of reviewers. The score sheets from the designated pair of

reviewers were compared for each of the individual steps, errors, and critical errors and the number of agreements (Agreement = both reviewers documented that a steps, or error or critical error was observed, or both scored the metric was not observed) was tabulated. In addition, the number of disagreements in observed metrics (i.e., disagreements = one rater reported observing the metric but the second rater reported not observing it) was tabulated. The interrater reliability (IRR) for the metric scores was calculated according to the following formula: Agreements/(Agreements + Disagreements).[23] The acceptable IRR was defined as 0.80 or greater.

Statistical Power Calculations

Power calculation: the numbers needed in each arm were based on transfer of training (ToT) effects observed in previous studies of PBP simulation studies where ToT rates of 42–69% were observed.[24–31] In the current study, we therefore expected to observe a decrease in performance errors >40%. A two-tailed test, with n = 16 trainees in each group with an alpha of 5% (which corresponds to a 95% confidence interval) and beta error 10% (i.e., 1–0.1 = 0.9 β) would yield a statistical power of 95%.

RESULTS

Table 1 shows the end of training results for the 2 groups assessed with the binary metric checklist (i.e., PBP) and the GEARS assessments. The mean and standard deviation scores show that PBP Group completed slightly more procedure steps but this difference was not statistically significant (F (1, 34) = 0.114, $P = 0.738$). In contrast, the PBP group made 75% fewer procedure errors than the conventional group, which was statistically significant (F (1, 34) = 16.426, $P < 0.001$). The critical error rate in both groups was low, but the conventional group made 94% more critical errors than the PBP group. This difference was however not statistically significant (F (1, 34) = 1.308, $P = 0.261$), possibly due to the large variability in scores exhibited by both groups. All the error scores combined (i.e., errors + critical errors = total errors) showed that the PBP-trained group made 74% fewer total errors than the Conventional trained group which was statistically significant (F (1, 34) = 16.904, $P < 0.001$).

The mean GEARS scores for both groups are also shown in Table 1. The PBP-trained group scored better (i.e., 15%) than the conventional group. The observed difference was not as large a magnitude as for the binary metric checklist for error scores but was statistically significant (F (1, 33) = 4.944, $P = 0.033$).

The proficiency benchmark for training outcome was quantitatively defined based on the objectively (and blinded) scored performance of the very experienced robotic surgeons performing the ORSI vesico-uretheral anastomosis task on the chicken model for the construct validity study.[17] To successfully conclude training, trainees needed to complete all 5 procedure steps, make no more than 10 errors and '0' critical errors. Table 1 shows that 67% of the PBP group demonstrated the proficiency benchmark in comparison to 17% in the conventional group. This difference was statistically significant ($\chi^2 = 9.26$, $P = 0.001$).

We assessed the capacity of the 2 assessments to discriminate the status of surgical trainees who demonstrated the proficiency benchmark as discrimination thresholds were varied, that is, internal structure validity evidence.[20] These are shown in Figure 2 receiver operator characteristic (ROC) curves for checklist and GEARS assessments. Accuracy is measured by the area under the ROC curve. The AUC for the binary checklist was 0.971 (SE = 0.027, asymptotic significance < 0.000, asymptotic 95% CI 0.919–1.00) and for GEARS it was 0.85 (SE = 0.066, asymptotic significance < 0.001, asymptotic 95% CI 0.721–0.98). Also shown in Figure 2 are interpolations on specificity of sensitivity levels of 0.8 and 0.9, respectively, for both checklist and GEARS assessments.

As shown in Figure 2, a sensitivity of 0.8 for binary checklist assessments on correct identification of proficiency status or specificity of trainees was excellent (i.e., 1–0.048 = 0.952). Specificity level was the same for a sensitivity level of 0.9. A sensitivity level of 0.8 for the GEARS assessments was lower than the sensitivity for the binary checklists (1–0.22 = 0.78 specificity) and lower again for a sensitivity of 0.9 (1–0.31 = 0.69 specificity).

We also assessed the IRR of the 2 assessment systems. The distribution of IRR score levels for all trainees with the binary checklist and GEARS assessments are shown in Figure 3. The median IRR for the binary checklist was 0.85, and the mean was 0.87. The median IRR for the GEARS assessments was 0.5, and the mean was 0.38. All the binary checklist IRR scores for individual trainees were above the 0.8 level (range = 0.2). The IRR levels of the GEARS assessments demonstrated considerably greater variability than the binary checklists, and only 2 were above the 0.8 IRR level (range = 0.83). The difference between the IRR levels of the 2 assessment approaches were compared with the nonparametric Mann-Whitney *U* test and found to be statistically significant ($z = -6.87$; $P < 0.000$).

DISCUSSION

The results from this study show that the PBP-trained group performed significantly better than the conventionally trained group. Both the binary checklist metrics and GEARS assessments demonstrated significant differences between the objectively assessed performance of the 2 groups. The performance differences as assessed with the binary checklist metrics demonstrated a greater magnitude of differences in comparison to the GEARS assessment, that is, 74% versus 15%. The steps binary metric demonstrated the smallest difference, and critical errors metric demonstrated the largest difference. Both these differences were however not statistically significant. In contrast, the error and total errors scores demonstrated >70% differences

TABLE 1.

The Mean Performance of the PBP and Conventional Trained Surgeons on the Checklist Metrics as well at the Percentage in Each Group Demonstration the Proficiency Benchmark

	PBP-Trained		Conventional Trained				
	Mean	SD	Mean	SD	% Difference	F value	Probability
Procedure steps	4.88	0.47	4.83	0.51	1%	0.11	0.738
Procedure errors	8.61	4.2	15.11	5.3	75%	16.43	0.000
Procedure critical errors	0.17	0.38	0.33	0.49	94%	1.31	0.261
Total errors	8.88	4.3	15.44	5.4	74%	16.9	0.000
Demonstrated the PBP proficiency benchmark	67%		17%		50%	$\chi^2 = 9.26$	0.002
GEARS Scores	22.23	3.7	19.39	3.9	15%	4.94	0.033

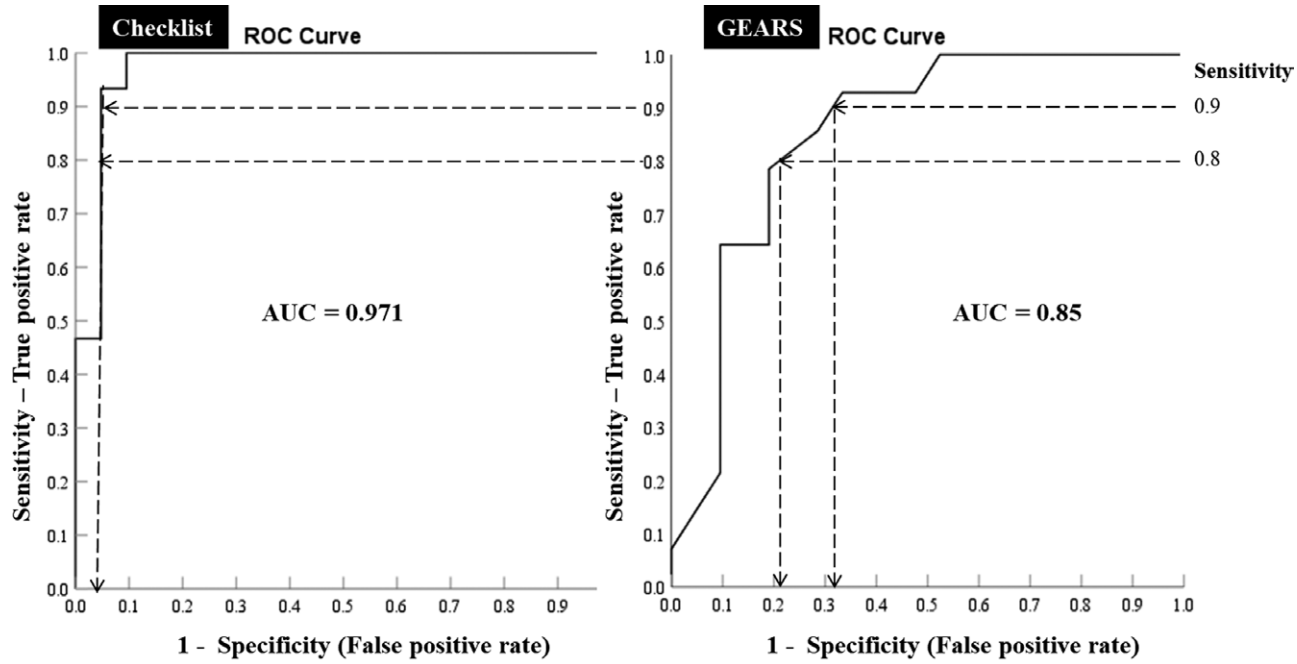

FIGURE 2. ROC showing the capacity of the checklist and GEARS assessments discrimination thresholds for surgical trainees who demonstrated the proficiency benchmark as discrimination thresholds were varied. ROC, receiver operator characteristic curves.

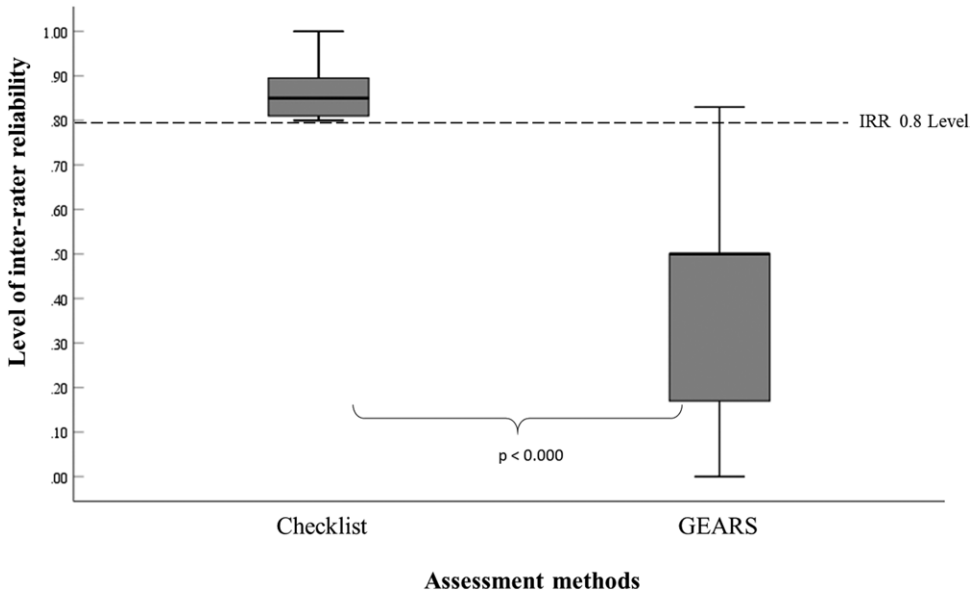

FIGURE 3. Boxplots showing the minimum value, the first quartile, the median, the third quartile, and the maximum value for the checklist and GEARS assessments of all trainees.

between the groups and were statistically significant. A more modest 15% difference between the groups was observed for the GEARS assessments, which also was statistically significant.

These large and statistically significant observed differences for the binary metrics probably originate from the more granular performance assessment that is integral to this assessment method. The binary metrics are primarily developed as tools for training,[27,28,32,33] which afford detailed, objective, and transparent formative performance feedback to the trainee.[8,14,34] This approach underpins a deliberate rather than repeated practice approach to training.[18] In contrast, the GEARS is a more global assessment tool, which does not necessarily provide detailed and explicit formative performance feedback to the trainee. The binary metrics are, however, task-specific and can only be applied for the specific task for which they were designed and are not designed to assess every basic robotic surgical skill. This offers the major advantage that performance feedback can be given to the trainee, which is explicit, objective, and transparent and has already been agreed by very experienced and skilled robotic surgeons and appears to accelerate quality assured learning by ~60% in comparison to quality assured conventional trainings.[35] Conversely, with GEARS, feedback is less specific, subjective, and more general.

This is the first study that we know which set out to compare the Sensitivity and Specificity of both approaches to the evaluation of basic robotic skills. In the context of this study, sensitivity refers to the ability of the assessment to correctly identify surgeons who had demonstrated the proficiency benchmark

by the end of training. Specificity refers to the capacity of the assessment to correctly identify surgical trainees who did not demonstrate the proficiency benchmark. The accuracy of the assessment method depends on how well the assessment separates the group being assessed into those who demonstrated the proficiency benchmark and those that did not demonstrate the benchmark. Accuracy is measured by the area under the ROC curve (AUC). An area of 1 represents a perfect assessment for predicting/identifying proficiency status correctly; an area of 0.5 represents an assessment method that is no better than chance at predicting proficiency status. A rough guide for classifying the accuracy of a diagnostic test is the traditional academic point system given below[36]:

- 0.90–1 = excellent
- 0.80–0.89 = good
- 0.70–0.79 = fair
- 0.60–0.69 = poor
- 0.50–0.59 = corresponds to a coin flip, that is, a useless model.

Both assessment methods used in this study were good or excellent when assessed by AUC. The binary checklist assessment method had excellent specificity at 0.8 and 0.9 levels of sensitivity. The specificity levels for the GEARS assessments were fair for a sensitivity level of 0.8 but poor for a sensitivity level of 0.9.

In terms of the most recent APA approach to validation, the task and standards used in this study fit well with a more unified approach to validation.[13] In the PBP methodology, the proficiency benchmark has always been based on the mean score of the objectively assessed performance of the experienced surgeons. This means it is objective, transparent, and fair, but probably more importantly, it is clinically meaningful and not an artificially defined performance level. Rather, it is based on the objectively assessed performance of experienced and practicing clinicians who are good at the surgical task. Thus, the surgical simulation task is derived from a part of a robotic prostatectomy that is integral to the procedure on a real patient. The procedure and task metrics have been agreed at a modified Delphi by very experienced surgeons who perform robotic prostatectomy[17,37] and both have demonstrated strong form and classic construct validity. Thus, the task has demonstrated evidence of content validity (i.e., Delphi consensus) on real patient surgeries and the simulation model used in this study. Furthermore, the passing benchmark is fair and transparent for trainees, as it is based on objectively assessed experienced surgeon performance. Performing the task poorly in a real patient will have significant consequences for the patient, that is, a anastomosis leak. It has also been demonstrated that simulation performance is highly correlated with real world performance (i.e., $r > 0.9$).[38] Additionally, prospective, randomized, and blinded clinical studies have demonstrated that (1) the vast majority of trainees demonstrate the proficiency benchmark with quality assured training and (2) that this translates into improved operating room/clinical performance[26,28–31,39] and improved patient outcomes.[25]

The data reported here on the psychometric assessment properties of the binary metrics and GEARS assessment tools support the notion of precision, effectiveness, and efficiency of binary performance metrics for training and assessment. Furthermore, in contrast to the Likert-type scale approach[40] to performance assessment a PBP-binary metrics approach has a tradition of >20 years of quantitative research[8,31,35] with a well established and articulated theory[8,34,41] that is supported by multiple sources of empirical data[24–26,28–31,35,38,39]—including the data presented here.

Possibly more important than the validity comments above are the reliability results on both assessment tools, which differed considerably. All the IRR scores for the Binary Checklist metrics were >0.8 but only 2 of the GEARS assessments reached this level. Overall, the IRR levels of the GEARS assessment were poor and this replicates previous findings on use of Likert scales in general[42,43] and GEARS specifically.[12] Given the widespread usage of these types of assessments, this is a worrying observation as a validated test that is demonstrated to be unreliable is by default not valid.[44] The findings from this study will require further scrutiny and investigation.

CONCLUSIONS

In this prospective, randomized, blinded controlled trial, it was shown that binary metrics for scoring a specific robotic VUA task demonstrated better levels of reliability, sensitivity, and specificity than the GEARS assessment.

ACKNOWLEDGMENTS

This study was conducted on behalf of the Junior ERUS/YAU working group on robot-assisted surgery of the European Association of Urology and the ERUS Education Working Group. EAU Robotic Urology Section (ERUS) Scientific Committee sponsorship of the 7 Belgian surgeons should be acknowledged.

REFERENCES

1. Dasgupta P. Robotics in urology. *Int J Med Robot Comput Assist Surg*. 2008;4:1–2.
2. Siegel R, Ma J, Zou Z, et al. Cancer statistics, 2014. *CA Cancer J Clin*. 2014;64:9–29.
3. Birkmeyer JD, Finks JF, O'Reilly A, et al; Michigan Bariatric Surgery Collaborative. Surgical skill and complication rates after bariatric surgery. *N Engl J Med*. 2013;369:1434–1442.
4. Sherbiny AE, Eissa A, Ghaith A, et al. Training in urological robotic surgery. Future perspectives. *Arch Esp Urol*. 2018;71:97–107.
5. Collins JW, Dell'Oglio P, Hung AJ, et al. The Importance of Technical and Non-technical Skills in Robotic Surgery Training [Figure presented]. *Eur Urol Focus*. 2018;4:674–676.
6. Puliatti S, Mazzone E, Dell'Oglio P. Training in robot-assisted surgery. *Curr Opin Urol*. 2020;30:65–72.
7. Sridhar AN, Briggs TP, Kelly JD, et al. Training in robotic surgery—an overview. *Curr Urol Rep*. 2017;18:58.
8. Gallagher AG, Ritter EM, Champion H, et al. Virtual reality simulation for the operating room: Proficiency-based training as a paradigm shift in surgical skills training. *Ann Surg*. 2005;241:364–372.
9. Chen J, Cheng N, Cacciamani G, et al. Objective Assessment of Robotic Surgical Technical Skill: A Systematic Review. *J Urol*. 2019;201:461–469.
10. Sánchez R, Rodríguez O, Rosciano J, et al. Robotic surgery training: construct validity of Global Evaluative Assessment of Robotic Skills (GEARS). *J Robot Surg*. 2016;10:227–231.
11. Ramos P, Montez J, Tripp A, et al. Face, content, construct and concurrent validity of dry laboratory exercises for robotic training using a global assessment tool. *BJU Int*. 2014;113:836–842.
12. Satava RM, Stefanidis D, Levy JS, et al. Proving the effectiveness of the fundamentals of robotic surgery (FRS) skills curriculum: a single-blinded, multispecialty, multi-institutional randomized control trial. *Ann Surg*. 2020;272:384–392.
13. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. American Educational Research Association, American Psychological Association & National Council on Measurement in Education. *Stand Educ Psychol Test*. 1999;2:9–24. Available at: https://www.bibguru.com/b/how-to-cite-standards-for-educational-and-psychological-testing/. Accessed April 17, 2023.
14. Gallagher AG. Metric-based simulation training to proficiency in medical education:- What it is and how to do it. *Ulster Med J*. 2012;81:107–113.
15. Gallagher AG, De Groote R, Paciotti M, et al. Proficiency-based progression training: a scientific approach to learning surgical skills. *Eur Urol*. 2022;81:394–395.
16. De Groote R, Puliatti S, Amato M, et al. Proficiency-based progression training for robotic surgery skills training: a randomized clinical trial. *BJU Int*. 2022;130:528–535.

17. Puliatti S, Mazzone E, Amato M, et al. Development and validation of the objective assessment of robotic suturing and knot tying skills for chicken anastomotic model. *Surg Endosc*. 2020;35.
18. Ericsson KA, Krampe RT, Tesch-Römer C. The role of deliberate practice in the acquisition of expert performance. *Psychol Rev*. 1993;100:363–406.
19. Pellegrini CA, de Santibañes E. Achieving Mastery in the Practice of Surgery. *Ann Surg*. 2019;270:735–737.
20. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. American Educational Research Association, American Psychological Association & National Council on Measurement in Education. *Stand Educ Psychol Test*. 1999;2:9–24. Available at: https://www.bibguru.com/b/how-to-cite-standards-for-educational-and-psychological-testing/. Accessed September 17, 2022.
21. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ*. 2003;37:830–837.
22. Cook DA, Brydges R, Ginsburg S, et al. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ*. 2015;49:560–575.
23. Kazdin AE. Parent management training: Evidence, outcomes, and issues. *J Am Acad Child Adolesc Psychiatry*. 1997;36:1349–1356.
24. Breen D, O'Brien S, McCarthy N, et al. Effect of a proficiency-based progression simulation programme on clinical communication for the deteriorating patient: a randomised controlled trial. *BMJ Open*. 2019;9:e025992.
25. Kallidaikurichi Srinivasan K, Gallagher A, O'Brien N, et al. Proficiency-based progression training: an end to end' model for decreasing error applied to achievement of effective epidural analgesia during labour: a randomised control study. *BMJ Open*. 2018;8:e020099.
26. Cates CU, Lönn L, Gallagher AG. Prospective, randomised and blinded comparison of proficiency-based progression full-physics virtual reality simulator training versus invasive vascular experience for learning carotid artery angiography by very experienced operators. *BMJ Simul Technol Enhanc Learn*. 2016;2:1–5.
27. Pedowitz RA, Nicandri GT, Angelo RL, et al. Objective assessment of knot-tying proficiency with the fundamentals of arthroscopic surgery training program workstation and knot tester. *Arthrosc J Arthrosc Relat Surg*. 2015;31:1872–1879.
28. Angelo RL, Ryu RKN, Pedowitz RA, et al. A proficiency-based progression training curriculum coupled with a model simulator results in the acquisition of a superior arthroscopic bankart skill set. *Arthrosc J Arthrosc Relat Surg*. 2015;31:1854–1871.
29. Van Sickle KR, Ritter EM, Baghai M, et al. Prospective, randomized, double-blind trial of curriculum-based training for intracorporeal suturing and knot tying. *J Am Coll Surg*. 2008;207:560–568.
30. Ahlberg G, Enochsson L, Gallagher AG, et al. Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. *Am J Surg*. 2007;193:797–804.
31. Seymour NE, Gallagher AG, Roman SA, et al. Virtual reality training improves operating room performance results of a randomized, double-blinded study. *Ann Surg*. 2002;236:458–63; discussion 463.
32. Angelo RL, Pedowitz RA, Ryu RKN, et al. The bankart performance metrics combined with a shoulder model simulator create a precise and accurate training tool for measuring surgeon skill. *Arthrosc J Arthrosc Relat Surg*. 2015;31:1639–1654.
33. Angelo RL, Ryu RKN, Pedowitz RA, et al. Metric development for an arthroscopic bankart procedure: assessment of face and content validity. *Arthrosc J Arthrosc Relat Surg*. 2015;31:1430–1440.
34. Gallagher AG, O'Sullivan GC, Gallagher AG, et al. Simulations for Procedural Training. In: *Fundamentals of Surgical Simulation*. Springer; 2011:39–66. doi:10.1007/978-0-85729-763-1_2
35. Mazzone E, Puliatti S, Amato M, et al. A systematic review and meta-analysis on the impact of proficiency-based progression simulation training on performance outcomes. *Ann Surg*. 2021;274:281–289.
36. The Area Under an ROC Curve. Available at: http://gim.unmc.edu/dxtests/roc3.htm. Accessed May 29, 2022.
37. Mottrie A, Mazzone E, Wiklund P, et al. *Objective assessment of intraoperative skills for robot-assisted radical prostatectomy (RARP): results from the ERUS Scientific and Educational Working Groups Metrics Initiative*. *BJU Int*. 2021;128:103–111.
38. Gallagher AG, Seymour NE, Jordan-Black JA, et al. Prospective, randomized assessment of transfer of training (ToT) and transfer effectiveness ratio (TER) of virtual reality simulation training for laparoscopic skill acquisition. *Ann Surg*. 2013;257:1025–1031.
39. Angelo RL, St Pierre P, Tauro J, et al. A proficiency-based progression simulation training curriculum to acquire the skills needed in performing arthroscopic bankart and rotator cuff repairs—implementation and impact. *Arthrosc J Arthrosc Relat Surg*. 2021;37:1099–1106.e5.
40. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84:273–278.
41. Gallagher AG. Metric-based simulation training to proficiency in medical education:- What it is and how to do it. *Ulster Med J*. 2012;81:107–113. Available at: https://pubmed.ncbi.nlm.nih.gov/23620606/. Accessed November 8, 2020.
42. 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;17:1525–1529.
43. Gallagher AG, O'Sullivan GC, Leonard G, et al. Objective structured assessment of technical skills and checklist scales reliability compared for high stakes assessments. *ANZ J Surg*. 2014;84:568–573.
44. Eignor DR. The standards for educational and psychological testing. *APA Handb Test Assess Psychol*. 2013;3:245–250.