Published in final edited form as:

Obstet Gynecol. 2014 June ; 123(6): 1193–1199. doi:10.1097/AOG.0000000000000288.

Validity and reliability of the robotic objective structured assessment of technical skills

Nazema Y. Siddiqui, MD, MHSc¹, Michael L. Galloway, Do², Elizabeth J. Geller, MD³, Isabel C. Green, MD⁴, Hye-Chun Hur, MD⁵, Kyle Langston, PA⁶, Michael C. Pitter, MD⁷, Megan E. Tarr, MD, MS⁸, and Martin A. Martino, MD⁶

¹Department of Obstetrics and Gynecology, Duke University, Durham, North Carolina

²Department of Obstetrics and Gynecology, Wright State University, Dayton, Ohio ³Department of Obstetrics and Gynecology, University of North Carolina, Chapel Hill, North Carolina ⁴Department of Obstetrics and Gynecology, Johns Hopkins University, Baltimore, Maryland ⁵Department of Obstetrics and Gynecology, Beth Israel Deaconess Medical Center, Boston, Massachusetts

⁶Department of Obstetrics and Gynecology, Lehigh Valley Health Network, Allentown, Pennsylvania ⁷Department of Obstetrics and Gynecology, Newark Beth Israel Medical Center, Newark, New Jersey ⁸Department of Obstetrics and Gynecology, Cleveland Clinic, Cleveland, Ohio

Abstract

Objective—Objective structured assessments of technical skills (OSATS) have been developed to measure the skill of surgical trainees. Our aim was to develop an OSATS specifically for trainees learning robotic surgery.

Study Design—This is a multi-institutional study in eight academic training programs. We created an assessment form to evaluate robotic surgical skill through five inanimate exercises. Obstetrics/gynecology, general surgery, and urology residents, fellows, and faculty completed five robotic exercises on a standard training model. Study sessions were recorded and randomly assigned to three blinded judges who scored performance using the assessment form. Construct validity was evaluated by comparing scores between participants with different levels of surgical experience; inter- and intra-rater reliability were also assessed.

Results—We evaluated 83 residents, 9 fellows, and 13 faculty, totaling 105 participants; 88 (84%) were from obstetrics/gynecology. Our assessment form demonstrated construct validity, with faculty and fellows performing significantly better than residents (mean scores: 89 ± 8 faculty; 74 ± 17 fellows; 59 ± 22 residents, $p < 0.01$). In addition, participants with more robotic console experience scored significantly higher than those with fewer prior console surgeries ($p < 0.01$). R-OSATS demonstrated good inter-rater reliability across all five drills (mean Cronbach's α : 0.79 ± 0.02). Intra-rater reliability was also high (mean Spearman's correlation: 0.91 ± 0.11).

Correspondence to: Nazema Y. Siddiqui, MD, MHSc, Duke University Medical Center, DUMC 3192, Durham, North Carolina 27710, Work: (919) 401-1006, Cell: (919) 599-9569, Fax: (919) 401-1033, nazema.siddiqui@duke.edu.

Presentation Information: Findings were presented at the 2013 APGO/CREOG Meeting, Phoenix, Arizona; Feb 27 – Mar 2, 2013

Conclusions—We developed an assessment form for robotic surgical skill that demonstrates construct validity, inter- and intra-rater reliability. When paired with standardized robotic skill drills this form may be useful to distinguish between levels of trainee performance.

Keywords

Surgical education; robotic surgery; simulation; technical skills; R-OSATS

Introduction

Robotic-assisted surgery has emerged as an alternative minimally invasive approach for gynecologic procedures including hysterectomy, myomectomy, cancer staging, and prolapse repair. The robotic approach offers increased dexterity with wristed laparoscopic instruments, but also requires technical training that is exclusive to robotic surgery.(1, 2)

Technical skills may be assessed using objective structured assessments of technical skills (OSATS).(3, 4) OSATS are typically standard assessment forms with pre-defined criteria indicating how to score performance on a technical skill. Compared to traditional surgical evaluations, OSATS allow for less biased assessments of technical performance, have demonstrated validity and reliability, and thus have been adopted broadly in obstetrics and gynecology(5-7) and other surgical training programs. (8-11) As surgical techniques have evolved, modified versions of the original OSATS form have been validated for assessment of laparoscopic and endoscopic procedures.(12)

Robotic surgical training relies heavily on simulation to develop technical skill with the instrumentation. Surgical trainees are often required to practice in a simulation environment prior to their involvement in live robotic procedures. However, when performing drills using an inanimate “dry-lab” simulator, it is difficult to know how to assess performance. Existing validated assessment tools such as OSATS(4) and the Global Operative Assessment of Laparoscopic Skills (GOALS)(12) assess parameters such as “Use of Assistants” and “Autonomy”, which are not relevant in robotic technical skills training. Thus, we further modified the OSATS and GOALS forms to be more useful for robotic simulation training, which we titled the Robotic-Objective Structured Assessment of Technical Skills (R-OSATS). Our primary aim was to assess the construct validity of this modified assessment form. Construct validity refers to the concept that a tool actually measures the construct that is being investigated.(13) In educational research, construct validity is often demonstrated by examining if an assessment tool can distinguish between levels of education (i.e. PGY year),(3, 6) which is the same paradigm we employed. We further aimed to assess inter- and intra-rater reliability of R-OSATS.

Materials and Methods

This study was approved by the Institutional Review Board at each participating center. Through consensus meetings, the investigators modified the OSATS form and developed the “Robotic-Objective Structured Assessment of Technical Skills” (R-OSATS), Appendix. R-OSATS is completed by directly observing performance on robotic simulation drills. Performance for each simulation drill is assessed across four categories: 1) depth perception/

accuracy; 2) force/tissue handling; 3) dexterity; and 4) efficiency. Each category is scored from 1-5, with higher scores indicating more proficiency. Scores are summed across categories, giving a maximum score of 20 per drill. Notably, we initially included a fifth category, called “instrument awareness”, that was designed to assess robotic arm collisions and whether the surgeon was moving an instrument outside the field of view. Some raters downgraded scores for large, “clinically relevant” collisions, whereas other raters incorporated every collision in their score. Thus, this category was removed after initial pilot testing, when we identified a large amount of variability in the interpretation among raters.

To test this modified assessment form, we created a standardized series of simulation exercises. These included five drills in sequence (Figure 1). The first drill was “tower transfer”, where the participant picks up rubber bands and transfers them to towers of varying heights. The second is “roller coaster”, where a rubber band is moved around a series of wire loops. The third drill is “big dipper” where a needle is placed into a sponge in various pre-specified directions. The fourth is “train tracks” which involves placing a running suture. The final drill is “figure-of-eight”, where the participant places a figure of eight suture and ties it using square knots. For the study, participating sites received a simulation training kit including all materials needed for the first two drills. For the needle/suturing drills (#3-5), sites were instructed to purchase 1-inch thick high-density foam, available at local craft stores. Foam was cut and marked using a template that was sent to each site.

We recruited residents, fellows, and faculty to participate in a multi-institutional study from January – May 2012. Participants were from gynecology (including gynecologic oncology, female pelvic medicine and reconstructive surgery, and advanced laparoscopy fellowships), urology, and general surgery. Participants attended study sessions where they completed baseline questionnaires assessing demographics, surgical experience, and prior robotic experience. They then reviewed an orientation video demonstrating the five robotic simulation drills, followed by completion of the drills. For each drill, participants were given one minute to practice, and permitted up to six minutes to complete the exercise. A different piece of foam was used for practice than for skill assessment. Drills were digitally recorded using the robotic camera. All materials (questionnaires and videos) were de-identified and labeled with study numbers. De-identified recordings were uploaded to a central data repository, which was electronically accessible by each investigator.

Each investigator was assigned recordings to score using the R-OSATS form. Prior to scoring, all investigators participated in an orientation to the assessment form and video review process. Each de-identified study recording was then randomly assigned to three blinded investigators. Notably, investigators were not assigned video recordings from their own site.

For our analysis, we first assessed construct validity by evaluating the ability of the assessment form to distinguish participants based on robotic surgical experience. This was evaluated by comparing scores based on *a*) level of training and *b*) reported number of prior robotic console surgeries. Using the three judges' R-OSATS scores, we calculated the median score for each exercise (maximum 20 points per exercise), and summed these

medians to calculate a summary score (maximum 100 points) for each participant. Three-way comparisons were performed with one-way analysis of variance (ANOVA), using the Tukey test for *post-hoc* comparisons. Two-way comparisons were performed using Student's t-test.

We also assessed inter- and intra-rater reliability. For inter-rater reliability, we compared the three judges scores for each drill using intraclass correlation coefficients (ICC), which are reported using Cronbach's α . To assess intra-rater reliability, multiple judges were re-assigned 10 of their original de-identified recordings for repeat review two months after the initial scoring process. Scores from the first and second review were compared using Spearman's correlation. Summarized intra-rater reliability was calculated using the means of the multiple judges' correlation coefficients. Correlation coefficients were compared between drills using ANOVA. Statistical comparisons of the Cronbach's alpha coefficients were performed with an extended Feldt's F-test using R version 3.0.1 (R Foundation for Statistical Computing, Vienna, Austria), and comparisons of the Spearman correlation coefficients were performed with an F test using SAS version 9.3 (SAS Institute, Cary, NC). The remaining data were analyzed using SPSS Version 20.0 (Chicago, Illinois) with $p < 0.05$ considered statistically significant.

Results

There were a total of 105 subjects, including 83 residents, 9 fellows, and 13 faculty members. The majority (84%) were from gynecology, with the remainder from general surgery and urology. Of the 83 residents, there were 13 PGY-1, 22 PGY-2, 29 PGY-3, 19 PGY-4, and no PGY-5 or 6. PGY-1 and 2 were considered junior residents while PGY-3 and 4 were considered senior residents. Of all residents, 10% had assisted at the console for more than 10 procedures. Fellows were in various stages of their training, with 5/9 (56%) having participated in more than 10 procedures as console surgeon. All faculty members had significant robotic console experience, with a median of 108 robotic procedures [range 50-500].

Construct validity was demonstrated by the ability of our assessment form to distinguish participants based on prior robotic surgical experience. Faculty performed significantly better than fellows, and fellows performed significantly better than residents (Table 2). Even amongst residents, our tool was able to distinguish between junior and senior level residents. Furthermore, participants with more robotic console experience scored significantly higher than those with less robotic console experience, regardless of PGY year (Table 3).

Inter-rater reliability was assessed for each drill. By convention, correlation coefficients above 0.7 are considered sufficiently high and coefficients above 0.8 are considered very high agreement.⁽¹⁴⁾ Inter-rater reliability was consistently high for all five drills, with a mean Cronbach's α of 0.79 ± 0.03 (Table 4). There were no significant differences in Cronbach's α coefficients between drills ($p=0.57$). Intra-rater reliability was also consistently high for all five drills, with a mean correlation coefficient (ρ) among all raters of 0.91 ± 0.11 (Table 5). The overall ANOVA indicates no significant differences in mean ρ between drills ($p=0.96$). There were also no differences in individual pairwise comparisons.

Discussion

We created a modified OSATS that can be used for objective assessment of performance during robotic training. This assessment form demonstrates construct validity, inter- and intra-rater reliability in inanimate simulation training.

There have been numerous recent reports focusing on robotic technical skills training. (15-17) To date, the more comprehensive studies have been performed using a series of nine inanimate technical skills - five of these based on prior Fundamentals of Laparoscopic Surgery (FLS) exercises, with four newly developed tasks for robotic surgery. (15, 18, 19) For these drills, investigators chose to assess performance using a combination of time and errors. Using this assessment technique, construct validity was established in a small study involving eight faculty/fellows and four medical students. (15) Thus, prior studies have shown that based on time and errors, trained surgeons perform better than medical students. We were less interested in the ability of an assessment technique to distinguish between novice and expert levels of proficiency, and more interested in the ability to distinguish performance among various levels of resident/fellow trainees. Furthermore, we wanted to offer a competency-based instead of time-based approach to assessment, which is currently lacking in robotic simulation training.(20) In our current study, we used a modified version of OSATS, as this method of assessment has demonstrated validity and reliability across numerous types of technical skills. (3, 4, 6, 12) Also, since our goal is to assess trainee performance during simulation training, we performed our validation study in a *resident and fellow trainee* population.

The strengths of our study lie in the methodology that we employed. Because of the multi-institutional design, we were able to include a larger number of resident and fellow trainees than typical for an educational study. We used rigorous methods for standardization of study sessions and blinding of scoring judges to reduce bias. Furthermore, our participants encompassed multiple surgical disciplines, and the judges represented many gynecologic subspecialties, allowing greater generalizability of our findings. Through our study design, we also demonstrated reproducibility of our results. This is important because, even if an educational tool is *valid*, it must also be *reliable* among different judges in order to be useful for a training program.

Our study also has some notable limitations. While there are multiple types of validity, we focused only on *construct* validity. We indeed demonstrated that our assessment form could discriminate between levels of training, and furthermore, between levels of prior robotic surgical experience. We did not however, include experienced non-robotic surgeons in our study. Inclusion of non-robotic faculty surgeons as a control group would have allowed us to assess for whether surgical experience alone, rather than robotic console experience, contributes to higher scores. Another limitation of our study is that we were unable to assess *predictive validity* - the ability of a tool to predict future performance (for example, in the live operating room). Lastly, although our intra-rater reliability coefficients were excellent, some of the inter-rater reliability scores were only moderately high (α of 0.78 for tower transfer, big dipper, and train tracks drills). This suggests that subtle differences in

interpretation may exist, though our ICC values were well within the 0.7 – 0.95 range reported in other OSATS studies.(11, 22-24)

Residency programs are continually being challenged to improve surgical training, and simulation training is becoming an increasingly popular method. For trainees wishing to learn robotic surgery, they often spend time practicing simulation drills, but lack an objective method to demonstrate their performance. We offer an assessment tool that can be used to assess trainees during simulation training. It is important to note that in order to use these drills and assessment form as a “test of readiness” for live surgery, rigorous methodology should be used to establish benchmark scores for resident performance. Importantly, for residents, who are still in a training environment, these benchmarks may be very different than those that are set for faculty surgeons. Furthermore, we recognize that robotic surgery not only involves technical console skills, but also presents unique challenges for the bedside assistant, communication in the operating room, and for cost containment in a training environment. Thus, we would consider incorporating technical skill training as one component of a comprehensive robotic training curriculum. One such curriculum exists and is currently undergoing further study.(25)

In summary, we present an assessment form for robotic simulation training that demonstrates validity and reliability in resident/fellow trainees. This tool, when paired with inanimate robotic skill drills, may prove useful for competency-based training and assessment.

Acknowledgments

Grace Fulton, BS (research coordinator), Duke University Medical Center, for study coordination. Samantha Thomas, MB, Duke University, for assistance with statistical analyses. Jon Kost & Hubert Huang, Lehigh Valley Health Network, for assistance with video editing and data coordination.

Disclosures: All authors received reimbursement for travel from Intuitive Surgical, Inc. for a total of three curriculum development/research meetings. Dr. Pitter is a paid consultant for Intuitive Surgical, Inc.; Dr. Geller has received speaking honoraria. Intuitive Surgical, Inc. did not sponsor or influence the study design, interpretation, or analysis of results.

Financial Support: Research reported in this publication was supported by the National Center For Advancing Translational Sciences under award number UL1TR001117. Dr. Siddiqui is supported by award number K12-DK100024 from the National Institute of Diabetes and Digestive and Kidney Diseases. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix

Robotic objective structured assessment of technical skills (R-OSATS). Assessment form and Lichert scale with scoring anchors.

References

1. Lenihan JP Jr, Kovanda C, Seshadri-Kreaden U. What is the learning curve for robotic assisted gynecologic surgery? *Journal of minimally invasive gynecology*. 2008 Sep-Oct;15(5):589–94. [PubMed: 18722971]
2. Woelk JL, Casiano ER, Weaver AL, Gostout BS, Trabuco EC, Gebhart JB. The learning curve of robotic hysterectomy. *Obstetrics and gynecology*. 2013 Jan; 121(1):87–95. [PubMed: 23262932]

3. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *The British journal of surgery*. 1997 Feb; 84(2):273–8. [PubMed: 9052454]
4. Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative “bench station” examination. *The American Journal of Surgery*. 1997; 173(3):226–30.
5. Buerkle B, Rueter K, Hefler LA, Tempfer-Bentz EK, Tempfer CB. Objective Structured Assessment of Technical Skills (OSATS) evaluation of theoretical versus hands-on training of vaginal breech delivery management: a randomized trial. *European journal of obstetrics, gynecology, and reproductive biology*. 2013 Sep 25.
6. Goff BA, Nielsen PE, Lentz GM, Chow GE, Chalmers RW, Fenner D, et al. Surgical skills assessment: a blinded examination of obstetrics and gynecology residents. *American journal of obstetrics and gynecology*. 2002 Apr; 186(4):613–7. [PubMed: 11967481]
7. Rackow BW, Solnik MJ, Tu FF, Senapati S, Pozolo KE, Du H. Deliberate practice improves obstetrics and gynecology residents' hysteroscopy skills. *Journal of graduate medical education*. 2012 Sep; 4(3):329–34. [PubMed: 23997877]
8. Gelinas-Phaneuf N, Del Maestro RF. Surgical expertise in neurosurgery: integrating theory into practice. *Neurosurgery*. 2013 Oct; 73(Suppl 1):S30–8.
9. Steehler MK, Chu EE, Na H, Pfisterer MJ, Hesham HN, Malekzadeh S. Teaching and assessing endoscopic sinus surgery skills on a validated low-cost task trainer. *The Laryngoscope*. 2013 Apr; 123(4):841–4. [PubMed: 23288689]
10. Tsagakaki M, Choudhary A. Mersey deanery ophthalmology trainees' views of the objective assessment of surgical and technical skills (OSATS) workplace-based assessment tool. *Perspectives on medical education*. 2013 Feb; 2(1):21–7. [PubMed: 23670654]
11. Zevin B, Bonrath EM, Aggarwal R, Dedy NJ, Ahmed N, Grantcharov TP, et al. Development, feasibility, validity, and reliability of a scale for objective assessment of operative performance in laparoscopic gastric bypass surgery. *Journal of the American College of Surgeons*. 2013 May; 216(5):955–65 e8. quiz 1029–31, 33. [PubMed: 23490542]
12. Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondré K, Stanbridge D, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *The American journal of surgery*. 2005; 190(1):107–13.
13. Messick JM Jr. Postoperative pain management. *Masui The Japanese journal of anesthesiology*. 1995 Feb; 44(2):170–9. [PubMed: 7739087]
14. Gallagher AG, Ritter EM, Satava RM. Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. *Surgical endoscopy*. 2003 Oct; 17(10):1525–9. [PubMed: 14502403]
15. Dulan G, Rege RV, Hogg DC, Gilberg-Fisher KM, Arain NA, Tesfay ST, et al. Proficiency-based training for robotic surgery: construct validity, workload, and expert levels for nine inanimate exercises. *Surgical endoscopy*. 2012 Jun; 26(6):1516–21. [PubMed: 22350226]
16. Hung AJ, Jayaratna IS, Teruya K, Desai MM, Gill IS, Goh AC. Comparative assessment of three standardized robotic surgery training methods. *BJU international*. 2013 Oct; 112(6):864–71. [PubMed: 23470136]
17. Lyons C, Goldfarb D, Jones SL, Badhiwala N, Miles B, Link R, et al. Which skills really matter? proving face, content, and construct validity for a commercial robotic simulator. *Surgical endoscopy*. 2013 Jun; 27(6):2020–30. [PubMed: 23389060]
18. Arain NA, Dulan G, Hogg DC, Rege RV, Powers CE, Tesfay ST, et al. Comprehensive proficiency-based inanimate training for robotic surgery: reliability, feasibility, and educational benefit. *Surgical endoscopy*. 2012 Oct; 26(10):2740–5. [PubMed: 22538678]
19. Dulan G, Rege RV, Hogg DC, Gilberg-Fisher KM, Arain NA, Tesfay ST, et al. Developing a comprehensive, proficiency-based training program for robotic surgery. *Surgery*. 2012 Sep; 152(3):477–88. [PubMed: 22938907]
20. Schreuder HW, Wolswijk R, Zweemer RP, Schijven MP, Verheijen RH. Training and learning robotic surgery, time for a more structured approach: a systematic review. *BJOG : an international journal of obstetrics and gynaecology*. 2012 Jan; 119(2):137–49. [PubMed: 21981104]

21. Messick S. Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*. 1995; 14(4):5–8.
22. Jabbour N, Reihsen T, Payne NR, Finkelstein M, Sweet RM, Sidman JD. Validated assessment tools for pediatric airway endoscopy simulation. *Otolaryngology--head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery*. 2012 Dec; 147(6): 1131–5. [PubMed: 22951432]
23. Marglani O, Alherabi A, Al-Andejani T, Javer A, Al-Zalabani A, Chalmers A. Development of a tool for Global Rating of Endoscopic Surgical Skills (GRESS) for assessment of otolaryngology residents. *B-Ent*. 2012; 8(3):191–5. [PubMed: 23113382]
24. Nimmons GL, Chang KE, Funk GF, Shonka DC, Pagedar NA. Validation of a task-specific scoring system for a microvascular surgery simulation model. *The Laryngoscope*. 2012 Oct; 122(10):2164–8. [PubMed: 22961593]
25. Robotic Training Network. [cited 1/12/2014]; Available from: <http://www.robotictraining.org>

Fig 1.

Images of five robotic skill drills: a) Tower transfer – transfer of rubber band from inner small towers to outer graduated height towers; b) Roller coaster – manipulate rubber band around wire loop; c) Big dipper – place needle into sponge in various arcs through pre-specified dot pattern; d) Train tracks – place a running suture with needle entering and exiting through dots; e) Figure of eight – place a figure of eight stitch with needle entering and exiting through dots followed by a square knot. Detailed information regarding drills and set-up is available at: www.roboticstraining.org.

Table 1
Summary of participant characteristics

	Residents (n=83)	Fellows (n=9)	Faculty (n=13)	P value*
More than 5 yrs LS surgery	4 (5)	1 (11)	12 (92)	<0.001
Prior robotic surgeries: >10 total (bedside or console)	34 (41)	7 (78)	13 (100)	<0.001
Prior robotic surgeries: >10 (console surgeon)	8 (10)	5 (56)	13 (100)	<0.001
Any prior robotic simulation training (dry lab, animal, or computer)	54 (65)	9 (100)	11 (85)	0.045
Any previous practice with 5 study drills	40 (48)	5 (56)	6 (46)	0.90

LS = laparoscopic; Data are displayed as # (%)

*
Based on Chi-square

Table 2
R-OSATS scores by level of training

	n	R-OSATS Score	P value
PGY 1&2 Residents	35	49 ± 23	<0.01 *
PGY 3&4 Residents	48	65 ± 20	
All Residents	83	59 ± 22	<0.01 ^
Fellows	9	74 ± 17	
Faculty	13	89 ± 8	

Data are displayed as mean ± SD

* Student's *t*-test

^ ANOVA with Tukey post hoc test (all 2-way comparisons significant)

Table 3
R-OSATS scores by prior console experience

Prior Console Surgeries	n	R-OSATS Score	P value [*]
None	44	49 ± 22	<0.001
1-10	35	66 ± 16	
>10	26	87 ± 8	

Data are displayed as mean ± SD

* ANOVA with Tukey post hoc test (all 2-way comparisons significant)

Table 4
Inter-rater reliability for skill drills

Drill	Cronbach's α^*	95% CI
Tower transfer	0.79	0.71 – 0.86
Roller coaster	0.84	0.77 – 0.89
Big dipper	0.75	0.65 – 0.83
Train tracks	0.78	0.70 – 0.85
Figure of eight	0.81	0.72 – 0.87
Mean (all 5 drills)	0.79 $\pm$ 0.03	

*
Intraclass correlation coefficient

Table 5

Intra-rater reliability

Rater	Tower Transfer	Roller Coaster	Big Dipper	Railroad Tracks	Figure of Eight
1	0.83	0.87	0.83	0.80	0.86
2	0.98	0.98	0.97	0.98	0.98
3	0.99	0.98	0.98	0.98	0.96
4	0.99	0.93	0.99	0.99	0.99
5	0.98	0.97	0.99	0.97	0.97
6	0.96	0.99	0.99	0.99	0.99
7	0.99	0.99	0.85	0.99	0.84
Mean	0.96 ± 0.06	0.96 ± 0.04	0.94 ± 0.07	0.96 ± 0.07	0.94 ± 0.06

Correlation of drill scores per rater using Spearman's correlation coefficients (ρ). ρ coefficients are summarized as mean $\pm$ standard deviation.