Received: 28 March 2022 | Revised: 7 December 2022 | Accepted: 8 December 2022

DOI: 10.1002/rcs.2492

ORIGINAL ARTICLE

The International Journal of Medical Robotics and Computer Assisted Surgery

WILEY

Effects of automated skill assessment on robotic surgery training

Jeremy D. Brown[1] | Katherine J. Kuchenbecker[2]

[1]Department of Mechanical Engineering, Johns Hopkins University, Baltimore, Maryland, USA

[2]Haptic Intelligence Department, Max Planck Institute for Intelligent Systems, Stuttgart, Germany

Correspondence
Jeremy D. Brown, Department of Mechanical Engineering, Johns Hopkins University, Baltimore, MD, USA.
Email: jbrow262@jhu.edu

Funding information
Intuitive Surgical, Grant/Award Number: Technology Research Grant

Abstract

Background: Several automated skill-assessment approaches have been proposed for robotic surgery, but their utility is not well understood. This article investigates the effects of one machine-learning-based skill-assessment approach on psychomotor skill development in robotic surgery training.

Methods: $N = 29$ trainees (medical students and residents) with no robotic surgery experience performed five trials of inanimate peg transfer with an Intuitive Surgical da Vinci Standard robot. Half of the participants received no post-trial feedback. The other half received automatically calculated scores from five Global Evaluative Assessment of Robotic Skill domains post-trial.

Results: There were no significant differences between the groups regarding overall improvement or skill improvement rate. However, participants who received post-trial feedback rated their overall performance improvement significantly lower than participants who did not receive feedback.

Conclusions: These findings indicate that automated skill evaluation systems might improve trainee self-awareness but not accelerate early stage psychomotor skill development in robotic surgery training.

KEYWORDS
automated skill assessment, inanimate training, robotic surgery

1 | INTRODUCTION

Robot-assisted minimally invasive surgery (RMIS) is becoming the standard of care in many surgical specialities.[1–3] Surgical platforms like Intuitive's da Vinci robot have been around for almost 2 decades and have played a major role in shaping the RMIS landscape and popularising its use in both routine and non-routine procedures. As a result, robot-assisted surgery as a clinical practice and medical industry has grown at an exponential rate, resulting in numerous general purpose and specialised robots at various stages of the development and deployment pipeline. While the introduction of each robotic platform brings new features and innovations intended to improve surgical practice, these differentiating features present a challenge with respect to surgical training and credentialling.

Given the increasing constraints on resident work hours and emphasis on patient safety,[4–7] a significant portion of early psychomotor skill development in minimally invasive surgery occurs through simulation-based training.[8] For RMIS, simulation-based training is provided through online and hands-on modules developed directly by robot manufacturers, training equipment manufacturers, or hospital training centres, and they utilise a combination of virtual reality (VR),[5,9–12] inanimate,[9,13–15] and in-vivo and ex-vivo training tasks.[9,16–18] Unlike laparoscopic[19] and endoscopic surgery,[20] there presently exists no standardised and widely accepted training curriculum for any RMIS platform or RMIS procedure.

Many VR platforms such as the da Vinci Surgical Skills Simulator have been validated for assessment of surgical skill[21] and have even shown potential to transfer basic robotic skills from simulation to the OR.[22] However, there is also evidence to suggest that the skills developed in VR are not as robust as those developed in the real-world.[23] Thus, for RMIS, training on the real clinical robot through inanimate, ex-vivo, or in-vivo tasks is still considered the gold standard.[9,10] Unfortunately, skill assessment for any task performed on the clinical robot requires a human rater to observe the performance, often through video review, and provide written or oral feedback. Though helpful for learning, structured human grading can be subjective, time consuming, and cost ineffective (as most raters are practicing physicians). Additionally, assessments are limited to features of skill that can be visually observed. Yet, it has been demonstrated for RMIS and other MIS approaches that the manner in which the surgeon physically interacts with the surgical environment is an indicator of skill.[14,24–32]

For RMIS procedures in particular, the ability to measure skill using robotic instrument motion, force, and vibration has led to the development of a number of approaches that utilise pattern-recognition algorithms to automatically assess surgical skill.[14,27,32–34] Often, these automated approaches utilise existing structured assessment metrics such as the Global Evaluative Assessment of Robotic Skill (GEARS)[35] or the Objective Structured Assessment of Technical Skill[36] and are developed using ground-truth skill ratings produced by trained surgical skill raters or crowd-sourced methods.[37–41] Additionally, some approaches have generated novel metrics based on the specific nature of the data and training task.[32] The benefits of these automated approaches are that they capture aspects of surgical skill performance that visual observation alone may miss, and that they provide feedback to the trainee very quickly and without overtaxing human raters. In this way, these advancements could help bring to clinical robot training one of the long-standing benefits of VR-based training approaches. Despite their great potential, however, there is limited evidence on the efficacy of these automated assessment approaches in improving surgical skill.

In this manuscript, we present the findings from one such investigation. Utilising an automated assessment approach previously developed by our research group,[14] we assessed the impact of providing post-trial scores from the GEARS assessment tool to novice trainees performing the peg transfer training task on a da Vinci surgical robotic platform. *We hypothesised that participants receiving feedback would improve their skill at peg transfer faster than what would occur through natural learning*. In what follows, we describe the automated assessment algorithm and the experimental protocol used to investigate its utility, along with the experimental results and a discussion of their implications in the broader context of RMIS training.

2 | METHODS

2.1 | Participants

We tested $N = 29$ participants (19 male, 10 female, mean age 25 ± 2 years) from the following two training levels ($N = 24$) second-, third-, and fourth-year medical students in the Agnew Surgical Society at the University of Pennsylvania Perelman Medical School, and ($N = 5$) first- and second-year surgical residents in the University of Pennsylvania Health System. Novice participants were specifically chosen for this study because their lack of prior robotic experience allowed for the greatest improvements in task performance. Of our 29 participants, 16 had no familiarity with the da Vinci surgical system, while the other 13 reported having limited familiarity; the remaining unchosen options were 'moderate' and 'extensive' familiarity. In addition, no participant had ever performed or assisted in a robotic case at the surgeon's console. Participants were compensated with a $15 gift card to offset the cost of travelling to the study location. Even with this monetary incentive, we found that the main motivation for participation was to gain experience on a da Vinci robot. All study procedures were approved by the University of Pennsylvania Institutional Review Board under protocol #825651. Participants were pseudo-randomised into one of two groups to balance gender and training level. Participants in the control group received no feedback regarding their performance on the training task. Participants in the feedback group received feedback from our automated skill-assessment system after every repetition of the task, as described below.

2.2 | Experimental setup

The study took place on a da Vinci Standard surgical system augmented with our Smart Task Board (STB) data collection system. The STB records the physical interactions from the patient-side manipulators of the Intuitive da Vinci surgical system and uses this data along with time-based measures to predict skill in robotic peg transfer according to the GEARS validated assessment tool.[35] The STB consists of three three-axis broad-bandwidth accelerometers that clip on the two primary robotic arms and the robotic camera arm, a task platform containing a three-axis force sensor, a custom signal conditioning and data acquisition circuit, a video recorder for recording the video feed from the robotic camera, and a pedal and light strip to control data recording. The STB predicts GEARS scores using a regression-based machine-learning algorithm that receives features from the accelerometer signals, force-plate signals, and time measures. This algorithm was developed using peg transfer data from participants of various skill level. More detail of the STB system, as well as the machine-learning algorithm development and evaluation, can be found in Brown et al.[14]

Participants used the da Vinci to perform the peg transfer task. Six triangular objects are placed on the left side of a pegboard. The participant picks up each object with their left tool, transfers it midair to their right tool, and places the object on a peg on the right side of the board, as shown in Figure 1A. After transferring all six objects, the participant returns the objects to the pegs on the left side of the board by reversing the process. Participants were instructed to retrieve objects that fall on the task board with the tool from which it fell. If an object fell off the task board, participants were instructed not to try to retrieve it. Peg transfer was completed with two 8-mm-diameter EndoWrist Maryland Bipolar Forceps tools. This relatively simple

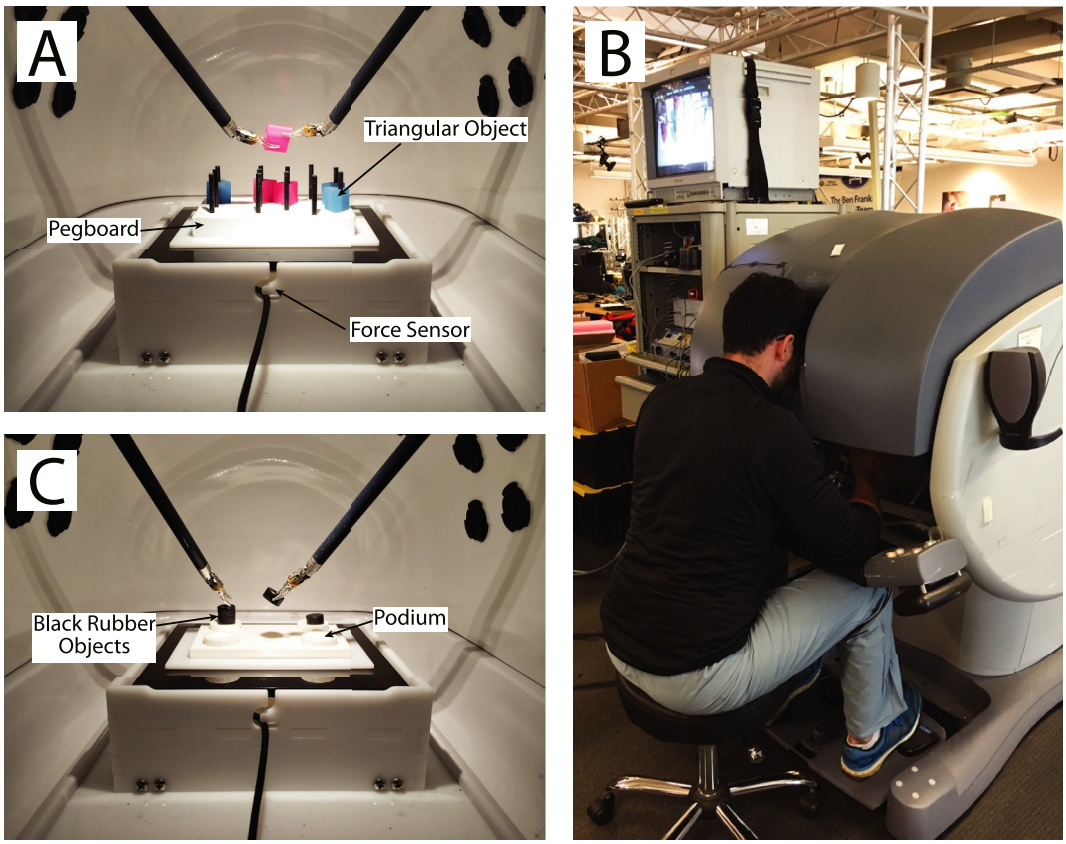

FIGURE 1 (A) Peg transfer task: Participants move the blue and pink triangular objects from the left side of the peg board to the right and then reverse the process. (B) A participant sits at the da Vinci surgeon's console to perform the peg transfer task. (C) Warm-up task featuring four black rubber objects that participants manipulated and moved from podium to podium to become familiar with the da Vinci operation.

task was adopted from the Fundamentals of Laparoscopic Surgery (FLS) manual skills test[42] and conducted in the da Vinci skills model shell to ensure consistent positioning of the camera and instruments (see Figure 1A).

2.3 | Experimental procedure

After giving informed consent, participants completed a demographic questionnaire. Each participant then sat at the da Vinci surgeon's console as shown in Figure 1B. The experimenter explained the da Vinci Standard system, including adjusting the ergonomics, focussing the camera, and clutching the tools and camera. Next, participants spent at least 5 min doing a warm-up task that featured four elevated podiums and four rubber objects that could be moved from one podium to another and stacked, as shown in Figure 1C.

After completing the practice session, the participant was shown how to operate the data/video recording system using the foot pedal and light strip. Participants then viewed static images depicting the peg transfer task procedure. Participants were subsequently given a written explanation of the five GEARS domains (Bimanual Dexterity, Depth Perception, Efficiency, Force Sensitivity, and Robotic Control) on which their performance would be evaluated. The GEARS domain Autonomy was not included in this study because the peg transfer task is simple enough that participants could complete it without verbal prompting. Although the descriptions of the GEARS domains are written in the context of skill evaluation in live surgery with real tissue, participants were instructed to interpret the language in the context of the inanimate peg transfer task. Participants were then shown the GEARS five-point evaluation survey shown in A and were instructed to perform the peg transfer task in an attempt to score as high as possible on each domain.

Before participants began the task, the tools were reloaded to reset their configuration, and the camera was adjusted to give a global view of the task board and the tool tips. Participants completed five trials of the peg transfer task. A short break of at least 2 min was taken in between trials. During this time, the tools and camera were reset, and the participant was shown the GEARS evaluation survey and instructed to think about ways of improving their score. Participants in the control group received no feedback about their performance. Participants in the feedback group, however, were shown performance feedback in the form of integer scores on the five GEARS domains mentioned above. These scores ranged from 1 (lowest) to 5 (highest) and were predicted by our automatic skill-evaluation system. The experimenter did not provide any explanation of the scores received, nor did he assist subjects in improving their scores. After completing all five trials of the peg transfer task, the participant completed a post-test survey that captured their

subjective assessment of the study. A separate survey was used for each of the two participant groups.

2.4 | Metrics and data analysis

To analyse any potential differences between our two participant groups, we have chosen as quantitative metrics the integral of the magnitude of the contact force vector, the trial duration, the root-mean-square (RMS) of the high-frequency (>100 Hz) and mid-frequency (20–100 Hz) accelerations of the left and right instruments, the raw GEARS scores for trials 1–5, and the overall learning rate for each of the five GEARS domains.

The force integral shows the total force the participant applied to the peg transfer task during a given trial. Raw GEARS scores were recorded for each domain and every trial for participants in both groups. These scores were computed using the regression-based algorithms discussed in the *Experimental Setup* section and detailed in Brown et al.[14] The learning rate for each domain was computed as the slope m of the line fitted to the GEARS scores received over all five trials using the formula $y = m \cdot x_t + b$: here y is the GEARS score for the selected domain, x_t is the trial number, and b is the y-axis intercept. An example of the linear fit for the Depth Perception GEARS scores for participant #4 is shown in Figure 2, where the resulting slope is $m = 0.4$ points per trial. Videos of the first and fifth trial for this participant are shown in https://youtu.be/yH0R-eXl3CQ and https://youtu.be/RsV6sm_obEw, respectively.

2.4.1 | Post-test survey

Our post-test survey represents a quantitative and qualitative assessment of each participant's opinions regarding the experiment. Participants in the control group were asked (1) how they felt their performance changed over the course of the experiment, (2) whether they would have preferred to receive feedback on their performance, (3) how feedback would impact their likelihood to practice, and (4) to rank each GEARS domain with respect to the amount of attention they gave it during the experiment. Participants in the feedback group were asked (1) how they felt their performance changed over the course of the experiment, (2) how useful it was to receive feedback on their performance, (3) how accurate they felt this feedback was, (4) how likely they would be to practice if they received feedback, (5) to rank each GEARS domain with respect to the amount of attention they gave it during the experiment, and (6) how the automatic evaluation system could be improved.

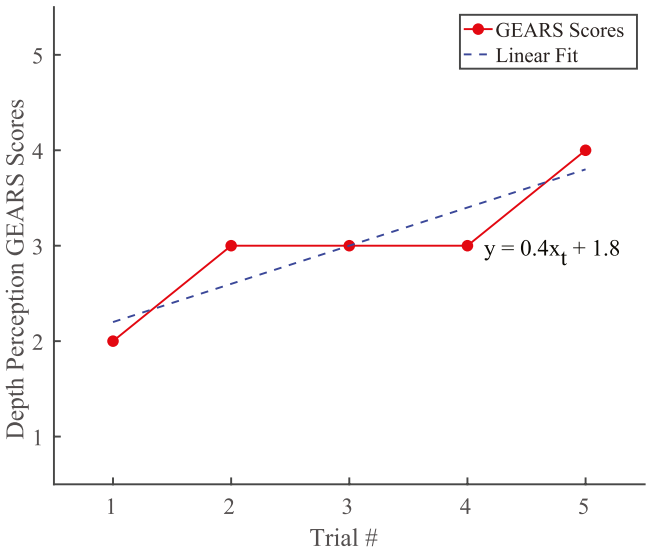

FIGURE 2 Linear fit of Depth Perception Global Evaluative Assessment of Robotic Skill (GEARS) scores for participant #4.

2.4.2 | Statistical analysis

All statistical analyses were performed using R (v.3.3.2). For all data, checks for normality and homogeneity of variance were performed using the Shapiro-Wilk and Levene Tests, respectively. Parametric t-tests and non-parametric Wilcoxon Rank sum tests were then used, where appropriate, to evaluate the force integral between groups for each trial, the trial duration between groups for each trial, the left and right tool accelerations between groups for each trial, the raw GEARS score differences between groups for each trial and each GEARS domain, including the overall GEARS score, the overall learning rate (slope) between groups for each GEARS domain and the overall GEARS score, and the qualitative response between groups to the survey question 'how did you performance change over the course of the experiment?'.

3 | RESULTS

3.1 | Quantitative

Our quantitative results suggest that for the peg transfer training task, receiving automated objective assessment in the form of scores on the GEARS assessment tool does not lead to faster skill development compared to natural learning. In particular, for our peg transfer training task, we found no significant difference in the average force magnitude integral (see Figure 3A) or the average trial duration (see Figure 3B) between our two groups in trials 1–5 ($p > 0.05$ for all comparisons). For the left and right tool accelerations, there was one significant difference in the mid-frequency right tool acceleration between the two groups for trial 1 ($p = 0.0133$), with the control group causing higher accelerations than the feedback group. However, this significant difference is not a result of our feedback system since it occurred on the first trial. All other differences were not significant ($p > 0.05$ for all comparisons) (see Figure 3C). We also found no significant difference for the average GEARS scores between our two groups in trials 1–5 for each of the five GEARS domains and the overall GEARS score ($p > 0.05$ for all comparisons) (see Figure 4A). Regarding the

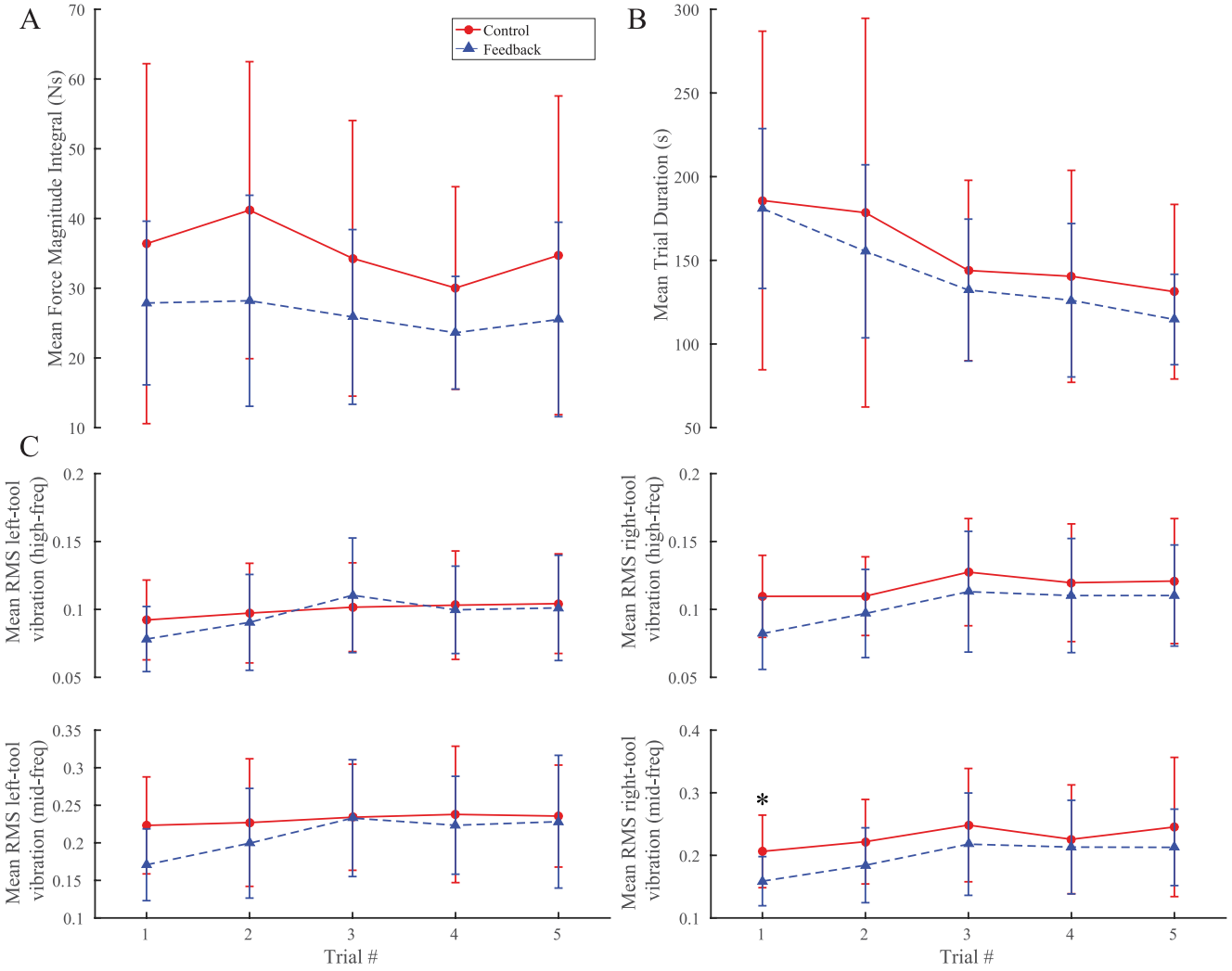

FIGURE 3 (A) Average force magnitude integral by participant group in trials 1–5. (B) Average trial duration by participant group in trials 1–5. (C) Average left and right tool acceleration (mid and high frequency) by participant group in trials 1–5. Solid red lines with circular markers refer to the control group. Dashed blue lines with triangular markers refer to the feedback group. Error bars represent ±1 standard deviation.

learning rate, we also saw no statistical difference between the two groups for any domain, as well as the overall learning rate, which is based on the overall score for each trial ($p > 0.05$ for all comparisons) (see Figure 4B).

3.2 | Qualitative

Participants in the feedback group (Median = 76 on a scale from 0 to 100) rated their overall improvement in performance significantly lower than participants in the control group (Median = 78) ($W = 132$, $p < 0.02$, $r = -0.48$), as shown in Figure 5. Additionally, participants in the control group gave a rating of 86 ± 12 (0–100 scale) as to how useful it would have been to receive feedback ratings. Participants in the feedback group gave an average rating of 76 ± 17 (0–100 scale) as to how useful it was to receive ratings. Participants in the feedback group gave an average rating of 63 ± 15 (0–100 scale) as to how accurate they felt the ratings were. About feedback influencing their desire to practice more, participants in the control group gave an average rating of 93 ± 14 (0–100 scale), and participants in the feedback group gave an average rating of 86 ± 12 (0–100 scale). Regarding participants' rated attention to each domain, we identified the most common (mode) rankings for each domain by group, as shown in Table 1.

4 | DISCUSSION

In this study, we sought to evaluate the utility of an automated skill-assessment platform for robotic surgery training. The assessment platform utilises measures of the physical interaction between the surgical robot and the surgical training environment in a regression-based algorithm to rate surgical skill according to the GEARS assessment survey. Novice participants were recruited to perform

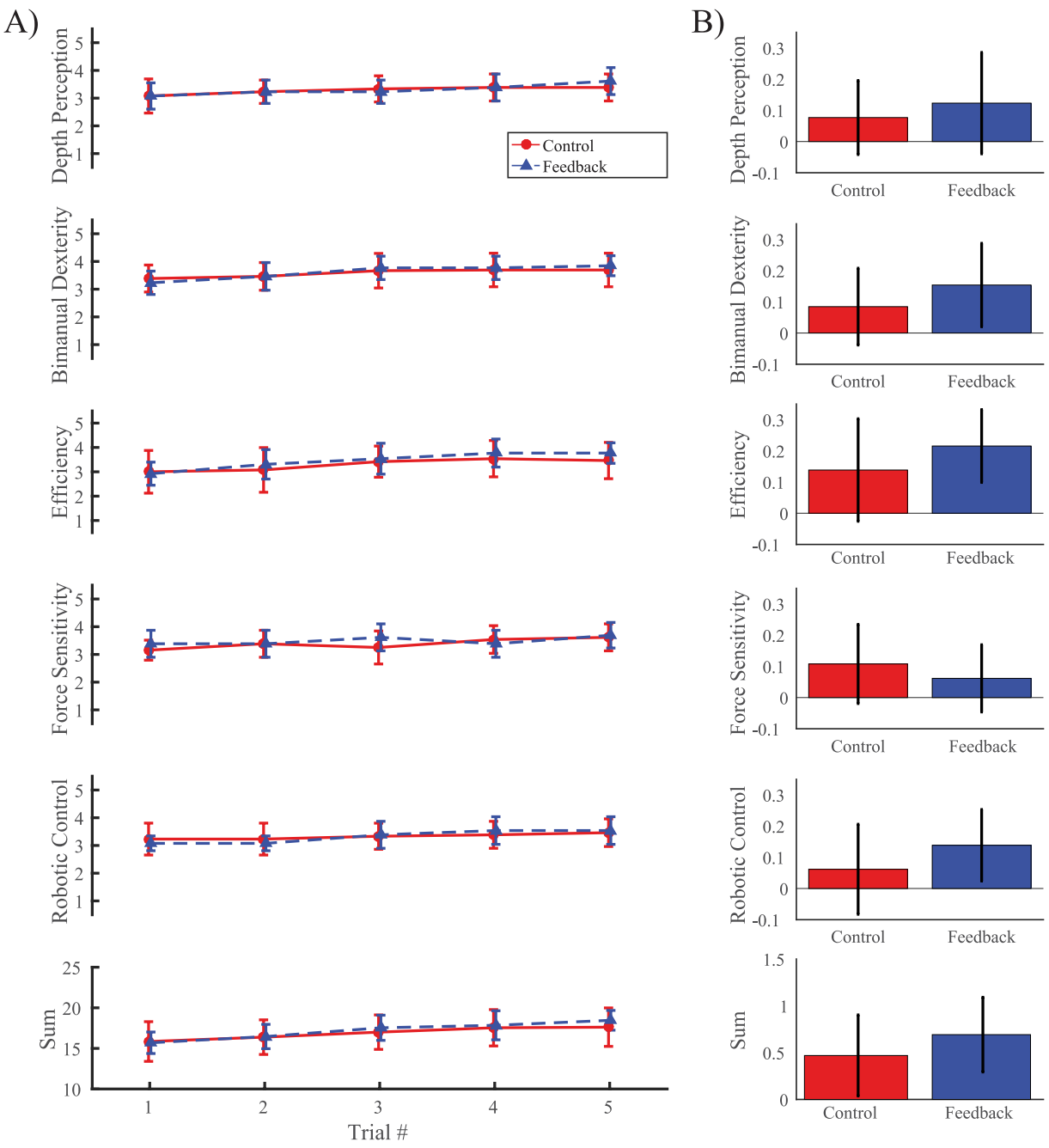

FIGURE 4 (A) Average Global Evaluative Assessment of Robotic Skill (GEARS) scores by participant group in trials 1–5 for each of the five GEARS domains and the overall GEARS score. Solid red lines with circular markers refer to the control group. Dashed blue lines with triangular markers refer to the feedback group. (B) Average learning rate by participant group for each of the five GEARS domains and the overall GEARS score. Error bars represent ±1 standard deviation.

the peg transfer psychomotor training task in two randomly assigned groups that differed in the availability of post-trial feedback of task performance. We found no quantitative difference between groups regarding their overall skill improvement throughout the training exercise. Qualitatively, however, we found that receiving post-trial feedback from our system affected participant's self-evaluation and motivations to practice, which could potentially play a more significant role for skill training involving more complicated tasks where natural learning occurs over a longer time.

Both groups of participants in this study improved their skill at the peg transfer task by the same amount. There were no significant differences between the force magnitude integral, trial duration, tool accelerations, and GEARS scores of the two groups after the first trial, indicating that the groups were well-balanced in the randomisation (see Figure 4). This balancing holds true for the remainder of the study (trials 2–5). That the group receiving feedback did not improve at a faster rate suggests that participants were not able to interpret and utilise the provided scores to make the necessary

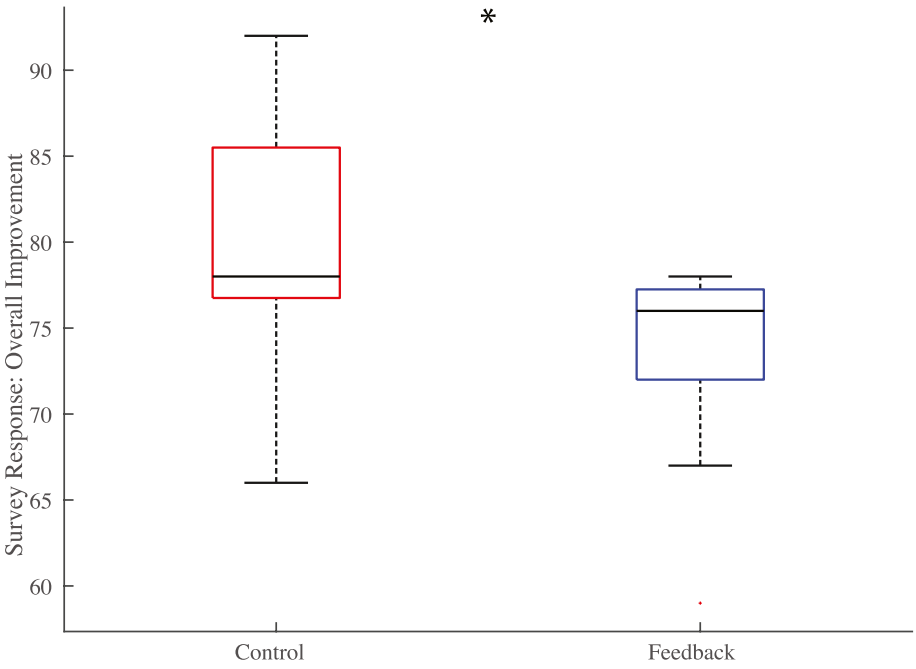

FIGURE 5 Survey response by participant group for the question 'How did your task performance change over the course of the study?'. The response scale ranged from 0 *'It got much worse'* to 100 *'It got much better'*.

TABLE 1 Participant rankings (scale of 1–5) of Global Evaluative Assessment of Robotic Skill (GEARS) domains with regard to attention paid to the domain during the experiment

GEARS domain	Control	Feedback
Depth perception	5	3
Bimanual dexterity	1	5
Efficiency	1	4
Force sensitivity	2	3
Robotic control	2	4

Note: Here 1 means '*Most Attention*', and 5 means '*Least Attention*'. Scores represent the mode (most common) ranking for participants in each group.

adjustments to improve their performance. It is worth reiterating here that the rubric that was given to participants for score interpretation used language that was written for performance evaluation on real tissue, not the inanimate peg transfer task, potentially making the task of score interpretation more difficult. Without guidance, participants were left to their own interpretations on how to best improve, which is not expected to differ significantly from participants in the control group, based on the group randomisation. This result, therefore, highlights the need not only for feedback, but also proper coaching, a common theme in the surgical training literature.[43,44]

It is possible that our short five-trial experiment captures only immediate skill improvement. On average, participants in both groups scored a three in all five domains (15 overall) on the first trial. This mid-range initial score limited the level of possible improvement to just two points (maximum score was five) for each domain. When we were developing the algorithm that produces the ratings, a score of five on any domain was, with few exceptions, obtainable only by expert robotic surgeons (>300 cases).[14] We would therefore reasonably expect that our non-expert participants in this short five-trial study to score no higher than a four on any domain, limiting the level of possible improvement even further to just a single point for each domain. How the scores between groups would change if the experiment were extended by another five or 10 trials is still unknown. Worth mentioning here is the fact that we also measure skill at a discrete integer level. Ratings on a continuous scale might show more between-subject and between-group variation.

Our findings also suggest that participants experienced the task differently in each group. The overall lower self-assessment by participants in the feedback group suggests these participants were influenced by the skill ratings they received as post-trial feedback. This effect of feedback is further supported by the fact that this difference in self-evaluation contradicts the lack of actual quantitative performance differences between the two groups. Taken another way, it appears that participants in the feedback group had a more realistic view of their performance, while participants in the control group had a more inflated view of their performance. Even with the lack of quantitative difference, it appears that receiving post-trial feedback allows for more accurate self-reflection on where one stands concerning skill proficiency.

Our qualitative findings also highlight that both groups thought it useful to receive ratings after every trial, and felt it would motivate them to practice more often. Of the findings presented in this manuscript, this one has potentially the most impact. Participants did not reject our system or view it as not being useful. Nor did the system make participants in the feedback group perform worse than

participants in the control group. The benefit of a system like this is that it alleviates the burden on a human rater to provide an evaluation of basic psychomotor skill development. It also provides ratings more efficiently; when developing our algorithm, we noticed that it took our human raters a minimum of 10 min to rate the same 10 trials that our automatic algorithm could rate in roughly 2 min. Another major benefit of an automated system is that it provides immediate feedback, as opposed to the long delay often associated with a human rater finding time to view a recorded video of the trial. Our algorithmic approach is also not affected by fatigue, conflicting responsibilities, or systematic bias as a human rater can often be.

When looking at the participants' assessment of the attention they paid to each of the GEARS domains, it becomes apparent that participants in the control group paid the most attention to the domains that were easiest to monitor visually or intuit. Participants could see both tools and could, therefore, monitor whether they were using both hands in equal proportions. The same may be true for Force Sensitivity and Robotic Control, where participants could visually estimate how well they were handling the task materials, and how smooth the overall operation of the robot was. Participants in this group probably paid the least amount of attention to Depth Perception because they lacked an objective way to measure it other than their knowledge that they were, in fact, perceiving depth. It is also worth noting here that the Efficiency GEARS domain aligns well with the metrics used in other surgical-skill-assessment evaluations, such as the FLS.

For participants in the feedback group, on the other hand, no one domain stood out. This lack of a universal preference suggests that participants were guided by the automatic GEARS ratings they were receiving, rather than only what they could visually discern during the task. Indeed, the post-trial feedback indicates that some participants were driven to pay attention to the domains in which they were doing the best, while others focused on domains where they were performing worst. With this current data set, neither participant strategy was dominant. It is also interesting to note that Efficiency and Bimanual Dexterity, which were ranked first most often by participants in the control group, were ranked fourth and fifth respectively most often by participants in the feedback group.

While our qualitative findings provide insight into the potential long-term benefits of an automated skill-assessment platform like the one presented here, the lack of quantitative differences between the two groups is a limitation that needs to be addressed before a system like this one can have a meaningful impact in surgical training. Of prime importance is investigating the impact of the system in a surgical training task with a significantly higher degree of difficulty. In this way, it can be reliably assumed that natural learning alone is insufficient to reach proficiency. Ideally, participants' initial GEARS scores would be in the 1–2 range on average instead of 3 as with the peg transfer task. More difficult tasks would also likely require more practice time to reach proficiency. Thus, any further investigations should consider more trial repetitions, including multi-session trials to understand the longitudinal effects of a system like this. In addition, future investigations could assess how the accuracy of skill feedback (e.g., placebo feedback) affects performance, and if a participant's performance would change after viewing other the performance and associated ratings of other participants. Future investigations could also consider whether trainee performance changes if they receive feedback on only one GEARS domain (e.g., force sensitivity) at a time, instead of all five domains together as in the current study.

Given that all of our participants were trainees, the times at which they were available for testing varied greatly. Thus, we had no control over participants' mental and physical fatigue levels prior to testing. While all participants appeared engaged in the experiment, future investigations should consider treating participants' fatigue and attention levels as covariates during analysis. Possible measurement approaches include reaction time testing or non-invasive neuroimaging approaches such as functional near-infrared spectroscopy.[45]

It may also be worth considering participants' prior exposure to and experience with other minimally invasive surgical techniques. Given the different training structures, modules, and rotations that exist in various residency programs, it cannot be assumed that all trainees are equivalent in skill proficiency. Thus, to understand the true benefits of a skill assessment platform like the one presented, future investigations should consider participants of all skill levels. Likewise, these investigations should recruit a large enough sample size to allow for robust sub-group analysis.

Finally, the use of automated assessment with a more difficult task may still fail to show any benefits. At present, we have sought to solve only one aspect of the surgical training process, that of assessment. Informing trainees of how 'good' or 'bad' their performance was is not the same as providing instructions for improvement. This reality was born out in our current results. Coaching, therefore, will be critical to any training platform, automated or not. The research into automated coaching is even more sparse than automated assessment, leaving open the opportunity for significant improvement. Exciting possible automated coaching solutions include providing haptic feedback,[46] visually displaying force exertion on tissue,[47] and notifying the trainee if they are neglecting one hand during a bimanual task. Still, the findings in this work should not be overlooked, as we have demonstrated several potential benefits of an automated skill-assessment platform in robotic surgery.

ACKNOWLEDGEMENTS

We thank Yu-Cheng Lou Lin and Kwame Owusu for helping implement the presented system for automatically calculating GEARS scores from physical measurements. Partial financial support was received from an Intuitive Surgical Technology Research Grant.

CONFLICT OF INTEREST

The authors declare that they have no non-financial interests to disclose.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Jeremy D. Brown https://orcid.org/0000-0001-5586-455X
Katherine J. Kuchenbecker https://orcid.org/0000-0002-5004-0313

REFERENCES

1. Sheetz KH, Claflin J, Dimick JB. Trends in the adoption of robotic surgery for common surgical procedures. *JAMA Netw Open*. 2020;3(1):e1918911. https://doi.org/10.1001/jamanetworkopen.2019.18911
2. McGuinness LA, Prasad Rai B. Robotics in urology. *Ann R Coll Surg Eng*. 2018;100(6):45-54.
3. Gorphe P. A contemporary review of evidence for transoral robotic surgery in laryngeal cancer. *Front Oncol*. 2018;8:121. https://doi.org/10.3389/fonc.2018.00121
4. Barden CB, Specht MC, McCarter MD, Daly JM, Fahey TJ. Effects of limited work hours on surgical training. *J Am Coll Surg*. 2002;195(4):531-538. https://doi.org/10.1016/s1072-7515(02)01242-5
5. Lee JY, Mucksavage P, Kerbl DC, Huynh VB, Etafy M, McDougall EM. Validation study of a virtual reality robotic simulator—role as an assessment tool? *J Urol*. 2012;187(3):998-1002. https://doi.org/10.1016/j.juro.2011.10.160
6. Greensmith M, Cho J, Hargest R. Changes in surgical training opportunities in Britain and South Africa. *Int J Surg*. 2016;25:76-81. https://doi.org/10.1016/j.ijsu.2015.11.052
7. Lambert TW, Smith F, Goldacre MJ. The impact of the European Working Time Directive 10 years on: views of the UK medical graduates of 2002 surveyed in 2013–2014. *JRSM Open*. 2016;7(3):205427041663270. https://doi.org/10.1177/2054270416632703
8. Higgins M, Madan CR, Patel R. Deliberate practice in simulation-based surgical skills training: a scoping review. *J Surg Educ*. 2020;78(4):1328-1339. https://doi.org/10.1016/j.jsurg.2020.11.008
9. Hung AJ, Jayaratna IS, Teruya K, Desai MM, Gill IS, Goh AC. Comparative assessment of three standardized robotic surgery training methods. *BJU Int*. 2013;112(6):864-871. https://doi.org/10.1111/bju.12045
10. Korets R, Mues AC, Graversen Ja, et al. Validating the use of the Mimic dV-trainer for robotic surgery skill acquisition among urology residents. *J Urol*. 2011;78(6):1326-1330. https://doi.org/10.1016/j.urology.2011.07.1426
11. Hertz AM, George EI, Vaccaro CM, Brand TC. Head-to-head comparison of three virtual-reality robotic surgery simulators. *J Soc Laparoendosc Surg*. 2018;22(1):1-6. https://doi.org/10.4293/jsls.2017.00081
12. Hogg ME, Tam V, Zenati M, et al. Mastery-based virtual reality robotic simulation curriculum: the first step toward operative robotic proficiency. *J Surg Educ*. 2017;74(3):477-485. https://doi.org/10.1016/j.jsurg.2016.10.015
13. Tergas AI, Sheth SB, Green IC, Giuntoli RL, Winder AD, Fader AN. A pilot study of surgical training using a virtual robotic surgery simulator. *J Soc Laparoendosc Surg*. 2013;17(2):219-226. https://doi.org/10.4293/108680813x13654754535872
14. Brown JD, Brien CEO, Leung SC, Dumon KR, Lee DI, Kuchenbecker KJ. Using contact forces and robot arm accelerations to automatically rate surgeon skill at peg transfer. *IEEE (Inst Electr Electron Eng) Trans Biomed Eng*. 2017;64(9):2263-2275. https://doi.org/10.1109/tbme.2016.2634861
15. 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(2):384-392. https://doi.org/10.1097/sla.0000000000003220
16. Aghazadeh MA, Jayaratna IS, Hung AJ, et al. External validation of global evaluative assessment of robotic skills (GEARS). *Surg Endosc*. 2015;29(11):3261-3266. https://doi.org/10.1007/s00464-015-4070-8
17. Hung AJ, Patil MB, Zehnder P, et al. Concurrent and predictive validation of a novel robotic surgery simulator: a prospective, randomized study. *J Urol*. 2012;187(2):630-637. https://doi.org/10.1016/j.juro.2011.09.154
18. Bertolo R, Garisto J, Dagenais J, Sagalovich D, Kaouk JH. Single session of robotic human cadaver training: the immediate impact on urology residents in a teaching hospital. *J Laparoendosc Adv Surg Tech*. 2018;28(10):1157-1162. https://doi.org/10.1089/lap.2018.0109
19. Peters JH, Fried GM, Swanstrom LL, et al. Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. *Surgery*. 2004;135(1):21-27. https://doi.org/10.1016/s0039-6060(03)00156-9
20. Vassiliou MC, Dunkin BJ, Fried GM, et al. Fundamentals of endoscopic surgery: creation and validation of the hands-on test. *Surg Endosc*. 2014;28(3):704-711. https://doi.org/10.1007/s00464-013-3298-4
21. Noureldin YA, Stoica A, Kassouf W, Tanguay S, Bladou F, Andonian S. Incorporation of the da Vinci surgical skills simulator at urology objective structured clinical examinations (OSCEs): a pilot study. *Can J Urol*. 2016;23(1):8160-8166.
22. Almarzouq A, Hu J, Noureldin YA, et al. Are basic robotic surgical skills transferable from the simulator to the operating room? A randomized, prospective, educational study. *Can Urol Assoc J*. 2020;14(12):416. https://doi.org/10.5489/cuaj.6460
23. Caccianiga G, Mariani A, De Momi E, Cantarero G, Brown JD. An evaluation of inanimate and virtual reality training for psychomotor skill development in robot-assisted minimally invasive surgery. *IEEE Trans Med Robot Bionics*. 2020;2(2):118-129. https://doi.org/10.1109/tmrb.2020.2990692
24. Reiley CE, Lin HC, Yuh DD, Hager GD. Review of methods for objective surgical skill evaluation. *Surg Endosc*. 2011;25(2):356-366. https://doi.org/10.1007/s00464-010-1190-z
25. Fried GM, Feldman LS, Vassiliou MC, et al. Proving the value of simulation in laparoscopic surgery. *Ann Surg*. 2004;240(3):518-528. https://doi.org/10.1097/01.sla.0000136941.46529.56
26. Rosen J, Hannaford B, Richards CG, Sinanan MN. Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. *IEEE (Inst Electr Electron Eng) Trans Biomed Eng*. 2001;48(5):579-591. https://doi.org/10.1109/10.918597
27. Lin HC, Shafran I, Yuh D, Hager GD. Towards automatic skill evaluation: detection and segmentation of robot-assisted surgical motions. *Comput Aided Surg*. 2006;11(5):220-230. https://doi.org/10.3109/10929080600989189
28. Ahmidi N, Gao Y, Béjar B, et al. String motif-based description of tool motion for detecting skill and gestures in robotic surgery. *Proc. Med Image Comput Comput Assist Interv (MICCAI). Part I. LNCS 8149*. 2013;16:26-33.
29. Gomez ED, Aggarwal R, McMahan W, Bark K, Kuchenbecker KJ. Objective assessment of robotic surgical skill using instrument contact vibrations. *Surg Endosc*. 2015;30(4):1419-1431.
30. Yamauchi Y, Yamashita J, Morikawa O, et al. Surgical skill evaluation by force data for endoscopic sinus surgery training system. *Proc. Med Image Comput Comput Asist Interv*. 2002;2488:44-51.
31. Trejos AL, Patel RV, Malthaner RA, Schlachta CM. Development of force-based metrics for skills assessment in minimally invasive surgery. *Surg Endosc*. 2014;28(7):2106-2119. https://doi.org/10.1007/s00464-014-3442-9

32. Liang K, Xing Y, Li J, Wang S, Li A, Li J. Motion control skill assessment based on kinematic analysis of robotic end-effector movements. *Int J Med Robot Comput Assist Surg*. 2018;14(1):e1845. https://doi.org/10.1002/rcs.1845
33. Despinoy F, Bouget D, Forestier G, et al. Unsupervised trajectory segmentation for surgical gesture recognition in robotic training. *IEEE (Inst Electr Electron Eng) Trans Biomed Eng*. 2016;63(6):1280-1291. https://doi.org/10.1109/tbme.2015.2493100
34. Ahmidi N, Poddar P, Jones JD, et al. Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. *Int J Comput Assist Radiol Surg*. 2015;10(6):981-991. https://doi.org/10.1007/s11548-015-1194-1
35. Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. *J Urol*. 2012;187(1):247-252. https://doi.org/10.1016/j.juro.2011.09.032
36. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84(2):273-278. https://doi.org/10.1046/j.1365-2168.1997.02502.x
37. Chen C, White L, Kowalewski T, et al. Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. *J Surg Res*. 2014;187(1):65-71. https://doi.org/10.1016/j.jss.2013.09.024
38. Holst D, Kowalewski TM, White LW, et al. Crowd-sourced assessment of technical skills: differentiating animate surgical skill through the wisdom of crowds. *J Endourol*. 2015;29(10):1183-1188. https://doi.org/10.1089/end.2015.0104
39. Polin MR, Siddiqui NY, Comstock BA, et al. Crowdsourcing: a valid alternative to expert evaluation of robotic surgery skills. *Am J Obstet Gynecol*. 2016;215(5):644
40. White LW, Kowalewski TM, Dockter RL, Comstock B, Hannaford B, Lendvay TS. Crowd-sourced assessment of technical skill: a valid method for discriminating basic robotic surgery skills. *J Endourol*. 2015;29(11):1295-1301. https://doi.org/10.1089/end.2015.0191
41. Wang Z, Reed I, Fey AM. Toward intuitive teleoperation in surgery: human-centric evaluation of teleoperation algorithms for robotic needle steering. In: *Proc. IEEE International Conference on Robotics and Automation (ICRA)*; 2018:5799-5806.
42. Fundamentals of laparoscopic surgery. [Online]. http://www.flsprogram.org/
43. Gagnon LH, Abbasi N. Systematic review of randomized controlled trials on the role of coaching in surgery to improve learner outcomes. *Am J Surg*. 2018;216(1):140-146. https://doi.org/10.1016/j.amjsurg.2017.05.003
44. Bonrath EM, Dedy NJ, Gordon LE, Grantcharov TP. Comprehensive surgical coaching enhances surgical skill in the operating room: a randomized controlled trial. *Ann Surg*. 2015;262(2):205-212. https://doi.org/10.1097/sla.0000000000001214
45. Thomas N, Ung G, Ayaz H, Brown JD. Neurophysiological evaluation of haptic feedback for myoelectric prostheses. *IEEE Trans Hum Mach Syst*. 2021;51(3):253-264. https://doi.org/10.1109/thms.2021.3066856
46. Brown JD, Fernandez JN, Cohen SP, Kuchenbecker KJ. A wrist-squeezing force-feedback system for robotic surgery training. In: *Proc. 2017 IEEE World Haptics Conference (WHC), no. June*. IEEE; 2017:107-112.
47. Reiley CE, Akinbiyi T, Burschka D, Chang DC, Okamura AM, Yuh DD. Effects of visual force feedback on robot-assisted surgical task performance. *J Thorac Cardiovasc Surg*. 2008;135(1):196-202. https://doi.org/10.1016/j.jtcvs.2007.08.043

How to cite this article: Brown JD, Kuchenbecker KJ. Effects of automated skill assessment on robotic surgery training. *Int J Med Robot*. 2023;19(2):e2492. https://doi.org/10.1002/rcs.2492

APPENDIX

A | GEARS RATING

GEARS

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		Uses both hands, but does not optimize interaction between hands		Expertly uses both hands in a complementary way to provide best exposure

Efficiency:

1	2	3	4	5
Inefficient efforts; many uncertain movements; constantly changing focus or persisting without progress		Slow, but planned movements are reasonably organized		Confident, efficient and safe conduct, maintains focus on task, fluid progression

Force Sensitivity:

1	2	3	4	5
Rough moves, tears tissue, injures nearby structures, poor control, frequent suture breakage		Handles tissues reasonably well, minor trauma to adjacent tissue, rare suture breakage		Applies appropriate tension, negligible injury to adjacent structures, no suture breakage

Robotic Control:

1	2	3	4	5
Consistently does not optimize view, hand position, or repeated collisions even with guidance		View is sometimes not optimal. Occasionally needs to relocate arms. Occasional collisions and obstruction of assistant		Controls camera and hand position optimally and independently. Minimal collisions or obstruction of assistant