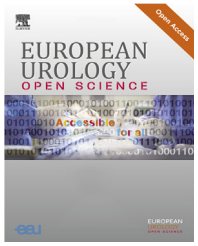
available at www.sciencedirect.com
journal homepage: www.eu-openscience.europeanurology.com



European Association of Urology

Education

Using Real-time Feedback To Improve Surgical Performance on a Robotic Tissue Dissection Task

Jasper A. Laca [a], *Rafal Kocielnik* [b], *Jessica H. Nguyen* [a], *Jonathan You* [a], *Ryan Tsang* [a], *Elyssa Y. Wong* [a], *Andrew Shtulman* [c], *Anima Anandkumar* [b], *Andrew J. Hung* [a,*]

[a] *Center for Robotic Simulation and Education, Catherine and Joseph Aresty Department of Urology, USC Institute of Urology, University of Southern California, Los Angeles, CA, USA;* [b] *Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA;* [c] *Thinking Lab, Department of Psychology, Occidental College, Los Angeles, CA, USA*

Article info

Article history:
Accepted September 26, 2022

Associate Editor:
M. Carmen Mir

Keywords:
Surgical education
Feedback
Robotic surgery
Learning
Mentoring

Abstract

Background: There is no standard for the feedback that an attending surgeon provides to a training surgeon, which may lead to variable outcomes in teaching cases.
Objective: To create and administer standardized feedback to medical students in an attempt to improve performance and learning.
Design, setting, and participants: A cohort of 45 medical students was recruited from a single medical school. Participants were randomly assigned to two groups. Both completed two rounds of a robotic surgical dissection task on a da Vinci Xi surgical system. The first round was the baseline assessment. In the second round, one group received feedback and the other served as the control (no feedback).
Outcome measurements and statistical analysis: Video from each round was retrospectively reviewed by four blinded raters and given a total error tally (primary outcome) and a technical skills score (Global Evaluative Assessment of Robotic Surgery [GEARS]). Generalized linear models were used for statistical modeling. According to their initial performance, each participant was categorized as either an innate performer or an underperformer, depending on whether their error tally was above or below the median.
Results and limitations: In round 2, the intervention group had a larger decrease in error rate than the control group, with a risk ratio (RR) of 1.51 (95% confidence interval [CI] 1.07–2.14; $p = 0.02$). The intervention group also had a greater increase in GEARS score in comparison to the control group, with a mean group difference of 2.15 (95% CI 0.81–3.49; $p < 0.01$). The interaction effect between innate performers versus underperformers and the intervention was statistically significant for the error rates, at $F(1,38) = 5.16$ ($p = 0.03$). Specifically, the intervention had a statistically significant effect on the error rate for underperformers (RR 2.23, 95% CI 1.37–3.62; $p < 0.01$) but not for innate performers (RR 1.03, 95% CI 0.63–1.68; $p = 0.91$).

* Corresponding author. University of Southern California Institute of Urology, 1441 Eastlake Avenue, Los Angeles, CA 90089, USA. Tel. +1 323 865 3700; Fax: +1 323 865 0120.
E-mail address: andrew.hung@med.usc.edu (A.J. Hung).

https://doi.org/10.1016/j.euros.2022.09.015
2666-1683/© 2022 The Author(s). Published by Elsevier B.V. on behalf of European Association of Urology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Conclusions: Real-time feedback improved performance globally compared to the control. The benefit of real-time feedback was stronger for underperformers than for trainees with innate skill.

Patient summary: We found that real-time feedback during a training task using a surgical robot improved the performance of trainees when the task was repeated. This feedback approach could help in training doctors in robotic surgery.

© 2022 The Author(s). Published by Elsevier B.V. on behalf of European Association of Urology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

A surgeon's formal training period involves learning from many mentors who provide feedback during surgery. The effectiveness of this feedback in improving the performance of a trainee surgeon ultimately dictates surgical outcomes. One of the main challenges with the current status quo is that there is no established standard for feedback delivered by mentors. One mentor's methodology may differ significantly from that of another. Trainee surgeons subjected to variable feedback may produce variable surgical outcomes, both during and after their formal training. This is compounded by the fact that some trainee surgeons are naturally gifted, while others need more help in their training [1].

Despite the challenges of the status quo, the benefits of surgical mentorship are hard to dispute. However, there is no means to continue mentorship beyond formal training, which represents another challenge. Young surgeons who have finished their formal training may lose out on the potential benefits of mentorship before they have achieved surgical mastery, a milestone that is likely to be achieved well after training.

Research has shown that surgical mentoring in robotic surgery can improve a trainee surgeon's task acquisition [2]. Without standardization, however, there is no guarantee that mentoring will consistently have a positive effect. Recent research has shown that the process of providing feedback can be further automated with an auto-mentor [3]. Automated feedback may solve the problem of mentor inconsistency and maintain the benefits of efficacious mentoring for as long as an individual might need it.

In a previous study we used feedback that was tailored to the individual trainee surgeon on the basis of their performance [4]. The summative feedback was provided weekly following each training session from the previous week. Individuals who received feedback had accelerated task acquisition in comparison to a control group with no feedback. While the feedback was able to improve the long-term performance, it was unable to improve performance of the immediate task.

In this present study we explored the effect of standardized real-time feedback on a simulated dissection task. The goal of the feedback is to aid the trainee surgeon's learning and to prevent or reduce errors during the training procedure.

We hypothesize that: (H1) feedback leads to improvements in surgical performance, measured as the error rate (H1a) and technical skills score (H1b), in comparison to a control; and (H2) participants who initially perform worse show a greater improvement after the intervention in comparison to those who initially perform well, measured as the error rate (H2a) and technical skills score (H2b).

2. Materials and methods

A group of novice medical students without any surgical experience completed a simulated dissection procedure on a da Vinci robotic system (Sunnyvale, CA, USA). The task involved: removal of the premarked top of a clementine skin and then exposing and removing a single segment of the interior fruit.

The study was designed as a two-task repetition spread over two separate sessions. The first session consisted of a brief training period, during which the participant was given a standardized introductory course on how to use the da Vinci surgical robot. This involved standardized instruction from a proctor, followed by two practice tasks. The practice tasks served as an opportunity for the participant to test all of the introductory skills needed to complete the experimental task.

The participant then completed a baseline/control clementine task (round 1). After all the participants had completed the first session, they were randomized into two groups (group 1 and group 2) with equal average performance scores. In the second session, participants completed the task once more (round 2), during which group 1 received feedback while group 2 did not. Depending on whether they scored above or below the median during round 1, participants in each group were individually categorized as either an innate performer (IP) or an underperformer (UP) for further analyses.

Endoscope video and audio (during the feedback round only) were recorded for the task in each round. Video was recorded by feeding the output of the endoscope into an external screen-grabber (OBS; https://obsproject.com/). Audio of the training sessions was recorded using Tobii eye tracking glasses (https://www.tobiipro.com/) and retrospectively synced with video of the experimental task.

Feedback consisted of seven prerecorded voice messages that were manually triggered by a proctor using an online soundboard (https://blerp.com). The seven pieces of feedback corresponded to seven risky behaviors commonly observed for novices, as identified from previous research on this dissection task [5]. Feedback was constructed according to the following standardized template: [warning/call to attention] + [risk factor] + [proposed mitigation] + [longer-term impact] [6]. Each piece of feedback was triggered by a specific risky behavior that could cause an error if not corrected in time. For example, if the participant did not appropriately grip the skin of the fruit, they were given feedback relevant to that behavior in an attempt to avoid the potential for a skin tear (error).

All feedback was recorded by the same voice actor (J.A.L.). Participants were instructed to stop what they were doing and listen to the entirety of the recorded feedback message before returning to the task. Each piece of feedback was preceded by an alert tone, which participants were told was their cue to stop and listen (Fig. 1).

Audio-free video of the tasks was retrospectively reviewed by four blinded raters. Each rater provided a total error tally and a technical skill score using the previously validated Global Evaluative Assessment of Robotic Surgery (GEARS) [7] for each video. Raters' intraclass correlation coefficients (ICCs) were measured to assess inter-rater reliability of scores for their first ten videos. Raters achieved ICC of 0.84 (95% confidence interval [CI] 0.783–0.892) for error identification and classification, and 0.81 (95% CI 0.742–0.862) for GEARS scores.

After completion of the feedback round, group 1 was asked to complete a System Usability Scale (SUS) questionnaire [8] to analyze the participants' perception of the feedback they received in terms of how useful it was.

2.1. Statistical analysis

Four hypotheses were tested in the study, two on the overall intervention effect (H1a and H1b) and two on interaction effects (H2a and H2b). To prevent inflation of the experiment-wise error rate (α) by multiple hypothesis testing, we used a fixed sequential method to assign the α value. In this procedure, we started with testing the overall group effect under the primary outcome. If the null hypothesis was rejected, the full experiment-wise error rate (α) was carried to the next test for the interaction between group and participant type. If the null hypothesis was rejected again, we moved to the secondary outcome (GEARS) and followed the same sequence to pass the experiment-wise error (α). Using this chain for hypothesis testing, if any test failed to reject the null hypothesis, we stopped the subsequent hypothesis testing and only reported the descriptive result. In an exploratory analysis, we used scatter plots to illustrate the pattern of correlation between GEARS, errors and SUS scores by the IP and UP participant types. Pearson or Spearman correlation was used, depending on data normality, for the descriptive analysis.

Generalized linear models were used for statistical modeling. We used a Gaussian distribution with an identity link function to model continuous outcomes (GEARS scores). The intervention effect is presented as the absolute difference between groups. For count outcomes (errors) we first standardized the measure by the total task length (number of errors/10 min) to derive error rates and modeled using a log link function; thus, the intervention effect is reported as a risk ratio (RR; multiplicative difference represented as the ratio of the error count/10 min between groups). The difference in intervention effect by participant types was tested using the interaction term in the model. The stratum-specific intervention effect was estimated using a post hoc contrast test. Model integrity was examined using residual plots and normality tests for residuals. Data analysis was performed using SPSS version 28.0.1.0 (SPSS Inc., Chicago, IL, USA).

3. Results

3.1. Participant demographics

A total of 45 medical students were recruited, of whom 13 were first-year and 32 were second-year students. The median age of the participants was 24 yr (range 21–34 yr) and 23/45 (51%) identified as female. There were no significant differences in demographic variables (gender, education, hand dominance, age, medical school year) between group 1 and group 2, or between the IP and UP groups.

In round 1, participants took between 4 min 47 s and 60 min 39 s to complete the task; the median time was 26 min 36 s (interquartile range [IQR] 31 min–32 min 23 s). In round 2, the median time was 15 min 3 s (IQR 11 min 44 s–22 min 59 s). A total of 60 instances of feedback were delivered to the intervention group (group 1). The median number of feedback instances was 3 (range 1–5). Among the five feedback categories (Table 1), the most frequently delivered was related to scissor usage, accounting for 28/60 instances (47%). The least common type of feedback was related to ideal tissue exposure and force (Table 1). Fisher's exact test revealed that there was no statistically significant difference in feedback type between the IP and UP groups (two-tailed $p = 0.369$). There was a statistically significant negative correlation between technical skills (GEARS score) and the error rate in both study rounds, with Spearman correlation coefficients of $r(41) = -0.69$ ($p < 0.001$) for round 1 and $r(41) = -0.83$ ($p < 0.001$) for round 2.

3.2. Intervention effect: change from baseline to round 2 between the intervention and control groups

We first estimated the impact of feedback delivery in comparison to the control (Fig. 2A). Comparison of the intervention (feedback) and control groups revealed a decrease in error rates in the feedback round, with a count RR of 1.51 (95% CI 1.07–2.14; $p = 0.021$). This meant that in the intervention group, the decrease in error rate was 1.51 times higher than in the control group (ie, repetition of the same task without feedback). Hypothesis H1a is supported. Similarly, we found an increase in technical skills score, with a mean difference between the groups of

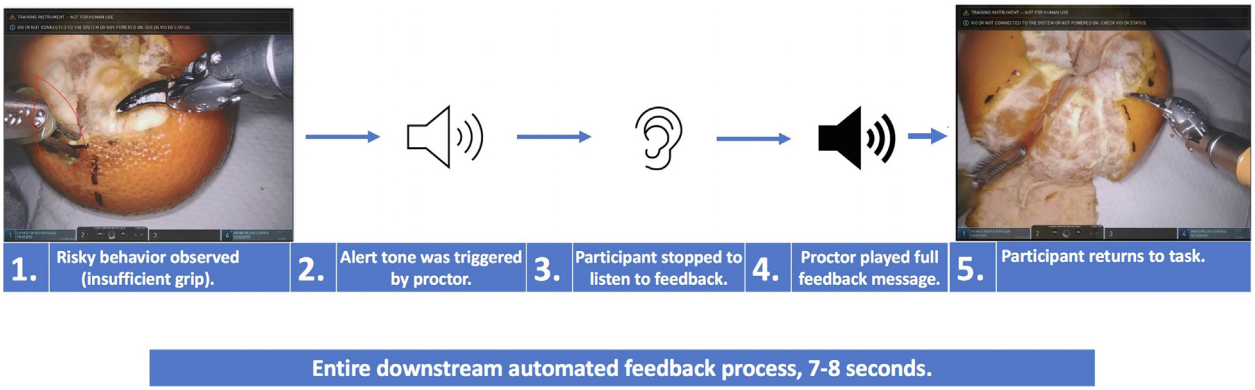


Fig. 1 – Example of the feedback delivery process.

Table 1 – Feedback delivery analysis: number of feedback instances by category during round 2 and the feedback phrases recorded for specific targeted behaviors

Feedback category	Targeted behavior	Instances, *n* (%)
Scissor usage	1. Misuse of scissors (peel)	15 (25)
	2. Misuse of scissors (fruit)	11 (18)
	3. Failure to follow lines	2 (3)
Retraction	4. Insufficient grip	22 (40)
Ideal surgical gesture	5. Disregard for surgical plane (peel)	4 (6)
Tissue exposure	6. Disregard for surgical plane (fruit)	3 (5)
Force sensitivity	7. Too much force	3 (5)
All		60 (100)
Specific feedback phrases for targeted behavior		
1. "Be mindful: avoid using the pointed end of the scissors when rotating the object, use the soft edge instead, this will reduce the risk of unintended skin punctures."		
2. "Be careful when using the scissors on the internal tissue, use the forceps to separate wedges to reduce the risk of unintentional punctures."		
3. "Be cautious: before attempting to remove any skin, make sure you have cut along the dotted lines. This will improve the efficiency of removing tissue."		
4. "Caution: if you're planning on pulling skin, make sure you are grabbing as much skin as you can to prevent it from tearing accidentally."		
5. "Remember: when removing the skin, use a peeling action after you have cut the lines; this will reduce unnecessary cutting and punctures."		
6. "Be aware: when you're removing the wedge, make sure you fully separate it from the rest of the fruit, otherwise it may come apart in pieces."		
7. "Be cautious with how much force you are applying with your tools; be as gentle as possible when appropriate, otherwise you risk damaging the object."		

2.15 (95% CI 0.81–3.49; p = 0.002; Table 2). This represents an increase in total GEARS score (scale 0–25) of more than 2 points. Hypothesis H1b is supported.

3.3. Intervention effect for the IP and UP groups

The interaction effect between performance group (IP vs UP) and condition (feedback vs no feedback) was statistically significant for the error rate, with $F(1,38) = 5.162$ (p = 0.029; hypothesis H2a is supported), but did not reach statistical significance for technical skills, with $F(1,38) = 1.588$ (p = 0.215; hypothesis H2b is rejected). For consistency, we performed a contrast test for both outcomes.

For the UP group there was a statistically significant difference between the intervention and the control in terms of a decrease in error rate, with RR of 2.23 (95% CI 1.37–3.62; p = 0.002). This indicates that UP participants reduced their error rate in the intervention by 2.23-fold in comparison to the control (task repetition without feedback). Similarly, for the UP group there was a statistically significant increase in technical skills score, with a mean difference of 2.99 (95% CI 0.90–5.06; p = 0.006; Table 2).

For the IP group the differences were not statistically significant: for the decrease in error rate the RR was 1.03 (95% CI 0.63–1.68; p = 0.914) and for the increase in technical skills score the mean difference was 1.32 (95% CI −0.36 to 2.99; p = 0.119; Table 2). To confirm that the IP and UP groups received a similar amount of feedback, we ran a Mann-Whitney U test to compare differences in the feedback instance counts. The test revealed no statistically significant (at α = 0.05) difference between the IP group (median 2.0, range 1–4; n = 10) and the UP group (median 3.0, range 2–5; n = 11), with U = 32.00 and z = −1.72 (two-tailed test, p = 0.087).

3.4. Usability of feedback

We further analyzed whether SUS scores were correlated with surgeon performance. For this analysis, we could only include participants who received feedback (group 1). A one-tailed Spearman correlation test between the decrease in error count and SUS score, controlling for baseline error counts (round 1), was statistically significant: $\rho(18) = 0.530$, p = 0.008. A one-tailed Pearson correlation test between the increase in technical skills and SUS score, controlling for baseline technical skills (round 1), was also statistically significant: $r(18) = 0.503$, p = 0.012.

We further assessed these correlations within the IP and UP subgroups separately (Fig. 3). For the UP group, usability was significantly correlated with a decrease in error rate: $\rho(8) = 0.674$, p = 0.016. Usability was also weakly significantly correlated with an increase in technical skills: $r(8) = 0.547$, p = 0.051. For the IP subgroup, usability was not significantly correlated with a decrease in error count: $\rho(7) = 0.162$, p = 0.339. In addition, usability was not significantly correlated with an increase in technical skills score: $r(7) = 0.123$, p = 0.376.

We further assessed whether the UP and IP groups differed in their average SUS score. The difference was not statistically significant, with a mean difference of 3.07 (95% CI −7.57 to 13.71; p = 0.572).

4. Discussion

Our study shows that real-time feedback led to a performance improvement (measured as technical skills and errors) for the simulated dissection task for which it was provided, and was particularly helpful for participants who initially struggled with the task (UP group). Furthermore, the performance of UP participants who received feedback was improved to a level similar to that of the IP participants. Broadly speaking, this serves as proof of concept that standardized, real-time surgical feedback can be a useful aid in surgical training.

SUS results were positively correlated with performance. The more receptive an individual was to the feedback (ie, how useful they found it), the more efficacious the feedback appeared to be. While both the UP and IP groups had statistically similar SUS scores, only the UP group had a statistically significant correlation between SUS score and performance. Regardless of how usable the feedback was perceived by the IP participants, they performed well. By contrast, the more usable the feedback was for UP participants, the better was their performance. One interpretation of this could be that the usability scores are a measure of how timely, understandable, and nondistracting the feedback was, but do not necessarily measure how "valuable" or "indispensable" the feedback was for accomplishing the

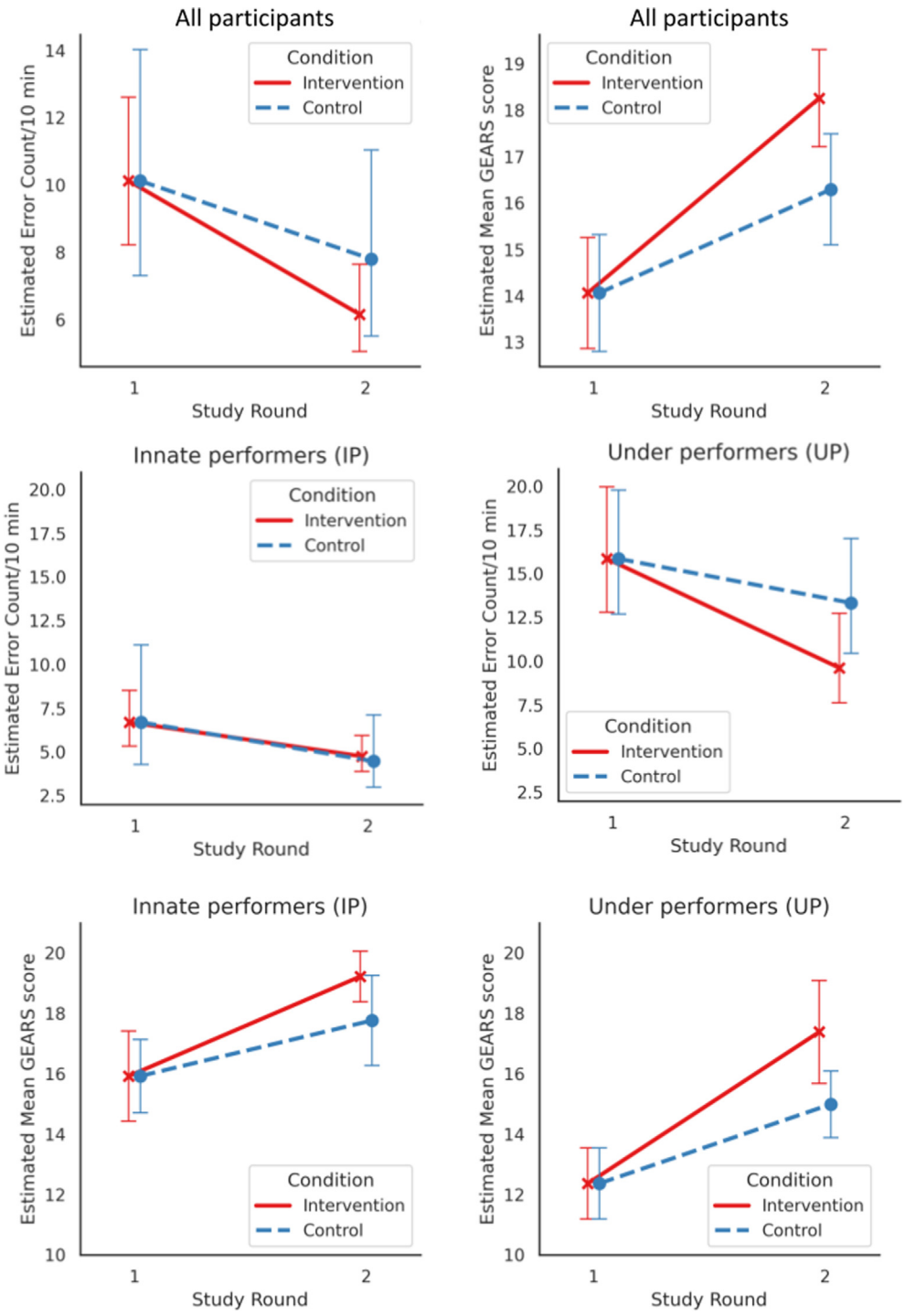


Fig. 2 – (A) Model-estimated mean error rate and technical skill score (GEARS) for rounds 1 and 2 for the feedback and control groups. (B) Model-estimated mean error rates for underperformers and innate performers. (C) Model estimated mean technical skill score (GEARS) for underperformers and innate performers. For all plots, error bars represent the 95% confidence interval for the estimates. GEARS = Global Evaluative Assessment of Robotic Surgery.

task. The IP participants may have understood the feedback, but it was not as useful or necessary for them. An alternative way to interpret these results is that the better someone does on a task, the more highly they will rate the intervention retrospectively.

One of the main limitations was the sample size and the representative population in our study. We had a limited pool of student participants who had no experience with a surgical robot. The 45 participants were introductory medical students, with no specialization towards a specific medical specialty. This was beneficial because the students represented a blank slate (without bias) at the beginning of the study. Their results can only be generalized to similar populations and not necessarily to surgeons at more advanced stages of training. Another limitation is the simulated task itself; while peeling and handling of a clementine

Table 2 – Impact of feedback on the error rate and technical skills in round 2 [a]

Group	Decrease in error rate [b]	Increase in technical skills [c]
	Count risk ratio (95% CI)	Mean difference (95% CI)
All participants (n = 43)	1.51 (1.07–2.14) *	2.15 (0.81–3.49) **
Underperformers (n = 22)	2.23 (1.37–3.62) **	2.99 (0.90–5.06) **
Innate performers (n = 21)	1.03 (0.63–1.68)	1.32 (−0.36 to 2.99)

CI = confidence interval.
[a] The analysis was performed for all 43 participants and an interaction contrast for underperformers and innate performers. Feedback had a significant impact on reducing the error rate and improving technical skills for underperformers, but not for innate performers. All models were controlled for performance at baseline (round 1).
[b] Higher values are better. Interaction p = 0.03.
[c] Higher values are better. Interaction p = 0.2.
* $p < 0.05$.
** $p < 0.01$.

is a relatively inexpensive and safe simulated model for tissue dissection, it cannot replicate live tissue dissection. Thus, our findings will have to be validated in future work with more life-like, complex tissue dissection tasks.

This study provides a step towards future automation of real-time feedback. However, because the feedback was not truly automated, our results are limited by the human element and do not represent results for an automated system. All the feedback that was presented during the study was triggered by a human proctor. To mitigate this variable, only one proctor was used throughout the study. Nonetheless, the conditions for provision of specific feedback were anchored to objective behaviors. Predefined feedback was given for specific behaviors. This was designed to reduce the feedback variability.

We propose that future studies should look to replicate our design for a variety of different surgical tasks and procedures. These additional studies should target participants of varying experience. Computer vision can be trained to recognize errors and behavioral patterns that lead to errors. These can then be used as cues for provision of targeted real-time feedback via an automated system.

5. Conclusions

The results indicate that our real-time feedback system was capable of improving surgeon performance in comparison to a control without feedback. Underperformers, in contrast to Innate Performers, benefited the most from the feedback, highlighting an ideal target audience for a future truly automated feedback system.

Author contributions: Andrew J. Hung had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Laca, Kocielnik, Nguyen, Hung.
Acquisition of data: Laca.
Analysis and interpretation of data: Kocielnik, Laca.
Drafting of the manuscript: Laca, Hung.
Critical revision of the manuscript for important intellectual content: Hung, Nguyen, Wong, You, Tsang, Shtulman.
Statistical analysis: Kocielnik.
Obtaining funding: Hung.
Administrative, technical, or material support: Hung, Nguyen, Laca.
Supervision: Anandkumar.
Other: None.

Financial disclosures: Andrew J. Hung certifies that all conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject matter or materials discussed in the manuscript (eg, employment/affiliation, grants or funding, consultancies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: Andrew J. Hung is a paid

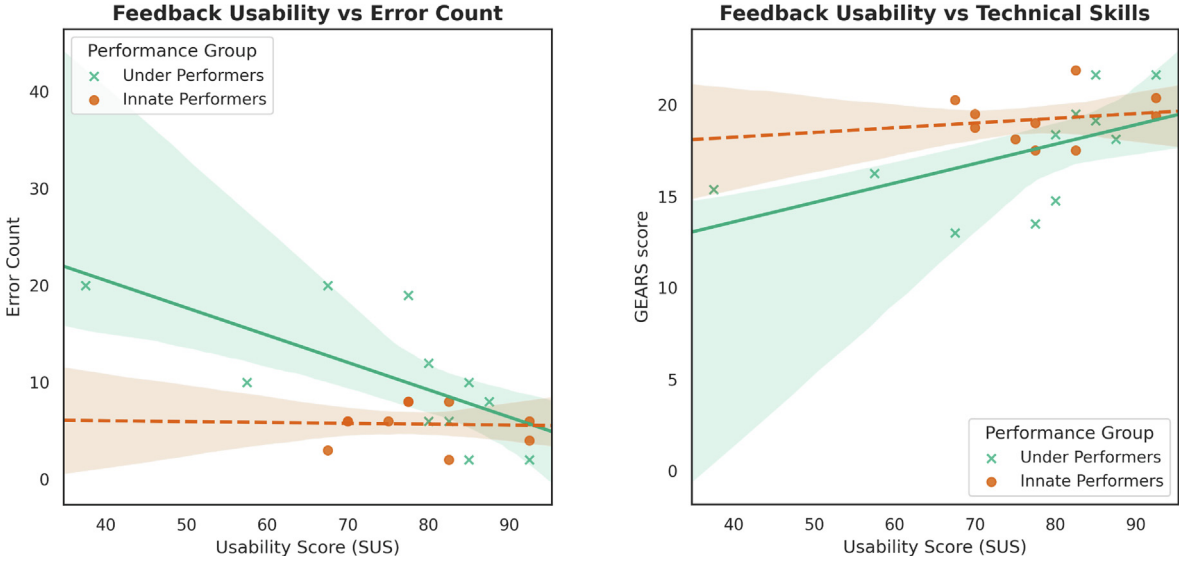


Fig. 3 – Empirical correlation of the System Usability Scale (SUS) score with the error count and technical skills score (GEARS) for underperformers and innate performers. The difference in slope shows that the feedback was more important in helping underperformers than innate performers. GEARS = Global Evaluative Assessment of Robotic Surgery.

consultant for Intuitive Surgical. Anima Anandkumar is a paid employee of Nvidia. The remaining authors have nothing to disclose.

***Funding/Support and role of the sponsor*:** This work was supported by the National Science Foundation under grant #2030859 to the Computing Research Association for the CIFellows Project. The sponsor played a role in analysis and interpretation of the data and review of the manuscript.

References

[1] El Boghdady M, Ewalds-Kvist BM. The innate aptitude's effect on the surgical task performance: a systematic review. Updates Surg 2021;73:2079–93. https://doi.org/10.1007/s13304-021-01173-6.
[2] Hanly EJ, Miller BE, Kumar R, et al. Mentoring console improves collaboration and teaching in surgical robotics. J Laparoendosc Adv Surg Tech 2006;16:445–51. https://doi.org/10.1089/lap.2006.16.445.
[3] Kopp KJ, Britt MA, Millis K, Graesser AC. Improving the efficiency of dialogue in tutoring. Learn Instruct 2012;22:320–30. https://doi.org/10.1016/j.learninstruc.2011.12.002.
[4] Ma R, Nguyen JH, Cowan A, et al. Using customized feedback to expedite the acquisition of robotic suturing skills. J Urol 2022;208:414–24.
[5] Nguyen JH, Chen J, Marshall SP, et al. Using objective robotic automated performance metrics and task-evoked pupillary response to distinguish surgeon expertise. World J Urol 2020;38:1599–605. https://doi.org/10.1007/s00345-019-02881-w.
[6] Duan P, Yocius M, Miltner M, Engler J, Schnell T, Uijt De Haag M. Human-in-the-loop evaluation of an information management and notification system to improve aircraft state awareness. In: Proceedings of the 2015 AIAA Infotech @ Aerospace Conference. Reston, VA: American Institute of Aeronautics and Astronautics; 2015. https://doi.org/10.2514/6.2015-0794.
[7] 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;187:247–52. https://doi.org/10.1016/j.juro.2011.09.032.
[8] Peres SC, Pham T, Phillips R. Validation of the System Usability Scale (SUS): SUS in the wild. Proc Hum Factors Ergonom Soc Annu Meet 2013;57:192–6. https://doi.org/10.1177/1541931213571043.