International Journal of Computer Assisted Radiology and Surgery (2020) 15:2101–2107
https://doi.org/10.1007/s11548-020-02253-5

ORIGINAL ARTICLE

Temporal variability of surgical technical skill perception in real robotic surgery

Jason D. Kelly[1] · Michael Nash[2] · Nicholas Heller[3] · Thomas S. Lendvay[4] · Timothy M. Kowalewski[1]

Received: 29 April 2020 / Accepted: 19 August 2020 / Published online: 29 August 2020

Abstract

Purpose Summary score metrics, either from crowds of non-experts, faculty surgeons or from automated performance metrics, have been trusted as the prevailing method of reporting surgeon technical skill. The aim of this paper is to learn whether there exist significant fluctuations in the technical skill assessments of a surgeon throughout long durations of surgical footage.

Methods A set of 12 videos of robotic surgery cases from common human patient robotic surgeries were used to evaluate the perceived technical skill at each individual minute of the surgical videos, which were originally 12–15 min in length. A linear mixed-effects model for each video was used to compare the ratings of each minute to those from every other minute in order to learn whether a change in scores over time can be detected and reliably measured apart from inter- and intrarater variation.

Results Modeling the change over time of the global evaluative assessment of robotic skills scores significantly contributed to the prediction models for 11 of the 12 surgeons. This demonstrates that measurable changes in technical skill occur over time during robotic surgery.

Conclusion The findings from this research raise questions about the optimal duration of footage needed to be evaluated to arrive at an accurate rating of surgical technical skill for longer procedures. This may imply non-negligible label noise for supervised machine learning approaches. In the future, it may be necessary to report a surgeon's skill variability in addition to their mean score to have proper knowledge of a surgeon's overall skill level.

Keywords Crowd sourcing · Surgical technical skill · Video segmentation · Bias

Introduction

Methods for assessing the technical skills of surgeons is paramount to ensuring to the public that surgeons are safe and effective. For years, summary scores or metrics have been used as the main method to report surgical skill. The most popular of these has been to use a Likert-scale scoring metric, from either non-expert crowds or from faculty surgeons. Past research has shown that crowds of non-experts concord with surgeon raters in evaluating technical skill [1]. Automated performance metrics have also been used, in which surgical events or streaming kinematic data have been used for computation of various metrics, suggesting superior objectivity [2]. As it is known that surgical skill is related to patient outcomes, and medical errors are the third leading cause of death in the USA, for which surgical errors contribute a large part, it remains important to accurately assess and report surgeon skill [3,4]. Much progress has been made by using statistical and machine learning models in the past with the goal of classifying surgeons into skill levels of 'novice' and 'expert' [5], but it remains a difficult computational task. For the largest statistical power, this usually leaves the most reliable approach being to obtain either surgeon or crowd evaluations from video [6].

✉ Jason D. Kelly
kell1917@umn.edu

1 Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN, USA

2 Department of Biostatistics, University of Washington, Seattle, WA, USA

3 Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA

4 Department of Urology, University of Washington, Seattle, WA, USA

One of the most popular robotic surgical skill assessment metrics is the Global Evaluative Assessment of Robotic Skills (GEARS), which is the most common objective assessment tool for robotic surgery [7] (Table 1). The subdomains in this metric include bimanual dexterity, efficiency, depth perception, force sensitivity and robotic control. Scores of 1–5 for each of these subdomains are totaled, for a cumulative score of 5–25, with higher scores belonging to more skilled surgeons. There is an additional domain in GEARS named Autonomy which is not used here. Autonomy is typically used when evaluating a surgeon's ability to work independently, which cannot be accurately assessed through video alone.

At least two confounding factors arise: (1) limitation of human attention span of the raters which inhibits reliability of their ratings and (2) the fluctuations of 'true technical skill,' naturally exhibited by a surgeon through time, when videos of a longer duration are used. It is unclear whether ratings remain reliable as durations scale up. The former is akin to *measurement noise* and is an attribute of the method to measuring skill—be it human ratings or computed computationally. The latter is akin to *process noise* and is an attribute of the surgeon's activities and tool–tissue environment interactions. Studies of human attention span have found that people viewing video lectures on average experienced significant decreases in attention and even stopped viewing after only about 6 min. One specific study found that including interactivity elements to video lectures led to an increase in watch time by at least 20% [8]. Given that evaluating surgeons requires an even higher degree of focus, using videos which are longer than 5 min may be detrimental to receiving accurate evaluations. Additionally, when making judgments, crowds can be susceptible to a contrast effect bias in which wide ranges in performance can lead to wildly different evaluations [9]. This could mean a performance of a novice who performed unusually well in the last segment of a video may receive unreliably high ratings. The natural fluctuations in technical skill remain largely unexplored, though hinted at in [10,11].

We hypothesize that statistically significant degrees of temporal fluctuations of perceived skill exist between smaller segments of surgical performances, which differ from previously obtained summary scores of the longer duration videos.

Methods

Dataset

This study utilized 12 videos from the Robotic Surgery Readiness (RSR) study [DoD TATRC Award Number W81XWH-15-2-0030]. The goal of that study was to determine whether preoperative warm-up on virtual reality tasks had a measurable improvement in surgical technical skills among practicing surgeons (no novice trainees) after typical periods of surgical inactivity, in a controlled randomized trial. This dataset contains 343 videos of surgical procedures consisting of live robotic surgeries, which were performed by 40 attending surgeons and trainees in urology, gynecology and general surgery at the University of Washington Medical Center and the Madigan Army Medical Center. These robotic surgeries were performed using da Vinci surgical robots, created by Intuitive Surgical (Sunnyvale, CA). Each video was manually edited to include roughly the first 15 min of surgical activity performed by the criterion surgeon. This beginning portion of the surgery contained similar actions among the different surgery specialties, such as suturing and cutting actions. A GEARS score for each 12–15-min performance was previously obtained from CSATS-Inc., using crowd evaluation.

The range of scores in all RSR videos was fairly small, with most lying between 20 and 22 out of 25 (only 15% of the full range possible). Our decision to obtain minute-by-minute ratings for a subset of videos rather than for all videos was motivated by our desire to obtain a sufficient number of ratings on each video segment with the limited resources available. This was the case as all surgeons assessed were practicing faculty and no trainees were involved in these videos. To obtain the largest possible range of skill from this dataset, six performances from the top quintile of scored performances and six from the bottom quintile of performances were used and given labels of 'expert' and 'proficient,' respectively, keeping in mind that all surgeons in this dataset are highly skilled, or they would not be allowed to operate in live robotic surgery procedures [12]. Additionally, half of each group of videos were performances in which the surgeon participated in a simulated surgery warm-up procedure beforehand, whereas the other half started surgery without warming up. This stratified formatting is illustrated in Table 2.

Crowd evaluation

Amazon Mechanical Turk was the crowd-sourcing platform used for this study, in which each non-expert crowd worker was paid $0.50 to watch and evaluate each of a series of 1-min videos. A Web domain was created for which Turkers would be redirected, where they submitted a consent form and were asked GEARS questions about videos. Each 12–15-min video performance was segmented into smaller videos having a duration of 1 min, using FFMPEG, an open source video editing tool [13]. Only 12 videos from the dataset were used due to limited monetary resources to compensate crowd workers with, while additionally having to receive multiple evaluations of each video to obtain reliable assessments.

Table 1 Likert-scale technical skill perception questionnaire, from the five domains of the GEARS assessment metric

Score	Depth perception
(1)	Constantly overshoots target, wide swings, slow to correct
(2)	
(3)	Some overshooting or missing of target, but quick to correct
(4)	
(5)	Accurately directs instruments in the correct plane to target
Score	**Bimanual dexterity**
(1)	Uses only one hand, ignores non-dominant hand, poor coordination
(2)	
(3)	Uses both hands, but does not optimize interaction between hands
(4)	
(5)	Expertly uses both hands in a complementary way to provide optimal exposure
Score	**Efficiency**
(1)	Inefficient efforts; many uncertain movements; constantly changing focus or persisting without progress
(2)	
(3)	Slow, but planned movements are reasonably organized
(4)	
(5)	Confident, efficient and safe conduct, maintains focus on task, fluid progression
Score	**Force sensitivity**
(1)	Rough moves, tears tissue, injures nearby structures, poor control, frequent suture breakage
(2)	
(3)	Handles tissue reasonably well, minor trauma to adjacent tissue, rare suture breakage
(4)	
(5)	Applies appropriate tension, negligible injury to adjacent structures, no suture breakage
Score	**Robotic control**
(1)	Consistently does not optimize view, hand position, or repeated collisions even with guidance
(2)	
(3)	View is sometimes not optimal. Occasionally needs to relocate arms. Occasional collisions and obstruction of assistant.
(4)	
(5)	Controls camera and hand position optimally and independently. Minimal collisions or obstruction of assistant.

Table 2 Summary information for 12 videos used from the RSR study

Group	Subgroup	Videos (*N*)	Prostatectomy	Hysterectomy	Other
Expert	Warm-up	3	1	1	1
	Control	3	1	0	2
Proficient	Warm-up	3	2	0	1
	Control	3	1	1	1

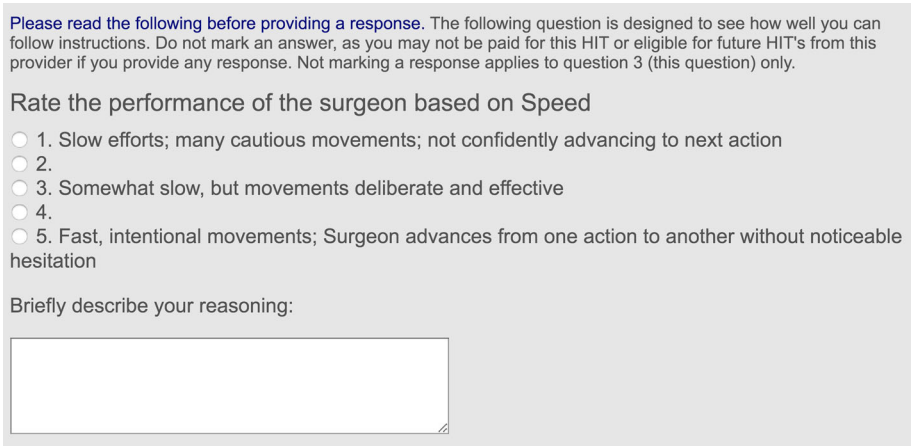

Fig. 1 Attention question inserted to the GEARS questionnaire, for quality assurance purposes. Note that skimming the instructions likely results in incorrect answers

The order of these videos was randomized so that crowds were viewing a non-chronological ordering. They were asked to provide a GEARS score for each minute-long video and proceeded to evaluate all remaining videos for the same performance. One-min segments were used as this provided a higher level of granularity to see differences in ratings throughout the performance. Smaller segments of videos, such as 30 seconds, appeared to be too short, and sometimes videos lacked meaningful surgical activity in these smaller segments. Forty raters were recruited to rate each performance.

A few extra elements were added to the user interface in an effort to increase interactivity and obtain better data quality. Once a crowd worker gave consent to participating in the study, they were shown a video with two performances, side by side, of a novice and expert performer of a laparoscopic surgical training task. The crowds were asked to evaluate which surgeon was better, and to what degree. After evaluating this task, the series of robotic surgery procedures began. After each GEARS Likert-scale question prompt, a text box was inserted, asking the raters to provide reasoning for their decision in each subdomain. These elements, such as the rater training portion of the study, have shown us in the past that we obtain more reliable ratings if these are included, as this might cause crowd workers to feel their ratings are more valued than when these elements are excluded.

Finally, an additional question which is not part of the standard GEARS questionnaire was inserted, asking for a rating of the 'speed' of the surgeon, shown in Fig. 1. Text in the prompt stated that the worker should not answer this question. This 'attention question' served as a mechanism to query which Turkers were focusing on the questionnaire versus those quickly finishing to be paid. All ratings which included answers to this 'attention question' were removed from analysis.

Statistical analysis

For each video, a linear mixed-effects model was estimated to analyze the effect of time on the outcome of individual GEARS ratings and how they changed from minute to minute. A linear mixed-effects model was chosen because it allowed us to estimate the expected score for each video segment and test the hypothesis that different segments of the same video had different expected scores while accounting for differences among raters. Each of the 12 models included a fixed effects term for each minute of video to estimate minute-to-minute differences in the expected score and a random intercept for each rater to account for differences between different raters. All statistical work was accomplished using R 3.6.2 [14], with data manipulation and visualization computed in Python 3.6 [15].

Results

Each video had 298 to 483 segment ratings made by 24 to 44 unique raters on 11 to 15 unique segments, and each segment was rated from 18 to 36 times, excluding those ratings dropped from analysis due to the rater's failure to correctly answer the 'attention question.' The mean of the GEARS scores at each minute of video for each expert and proficient surgeon is shown in Fig. 2. No identifiable relationship or trend in any of the four subplots is immediately apparent. However, the linear mixed-effects model did find that there were significant differences within each surgeon's scores as the videos progressed. For 11 of the 12 videos, modeling the change in GEARS scores over time significantly contributed to the prediction model, as compared with a 'null' model containing no terms for differences in scores between different segments of the same videos. For the first of the three videos showing a 'proficient' surgeon who did not receive the inter-

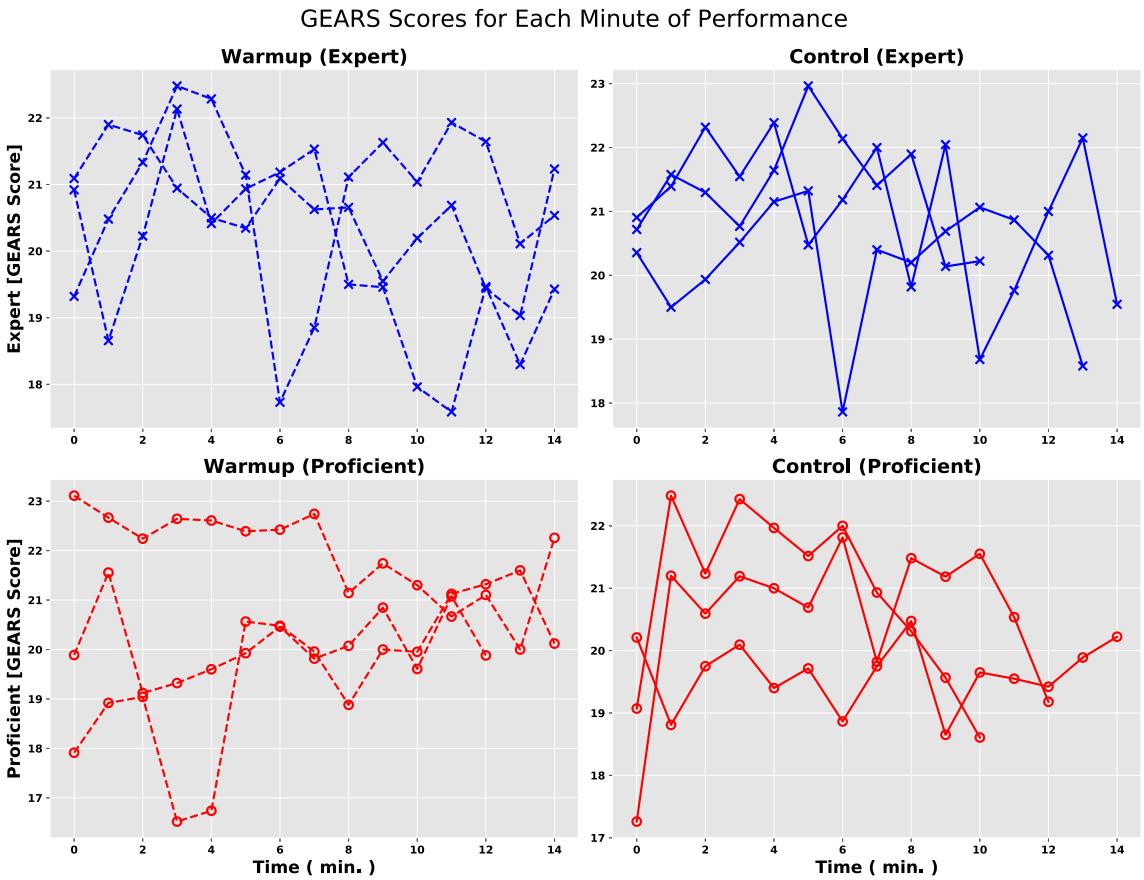

Fig. 2 All mean crowd evaluations from each proficient (red) and expert (blue) surgeon at each minute of the performance. Surgeons previously randomized to the control (solid lines) group normally started surgery without any intervention or preoperative warm-up used. Warm-up (dashed lines) group surgeons reviewed a virtual reality warm-up module prior to surgery (the warm-up hypothesis from the original randomized study is not being tested or evaluated in this research, only temporal variation in ratings is)

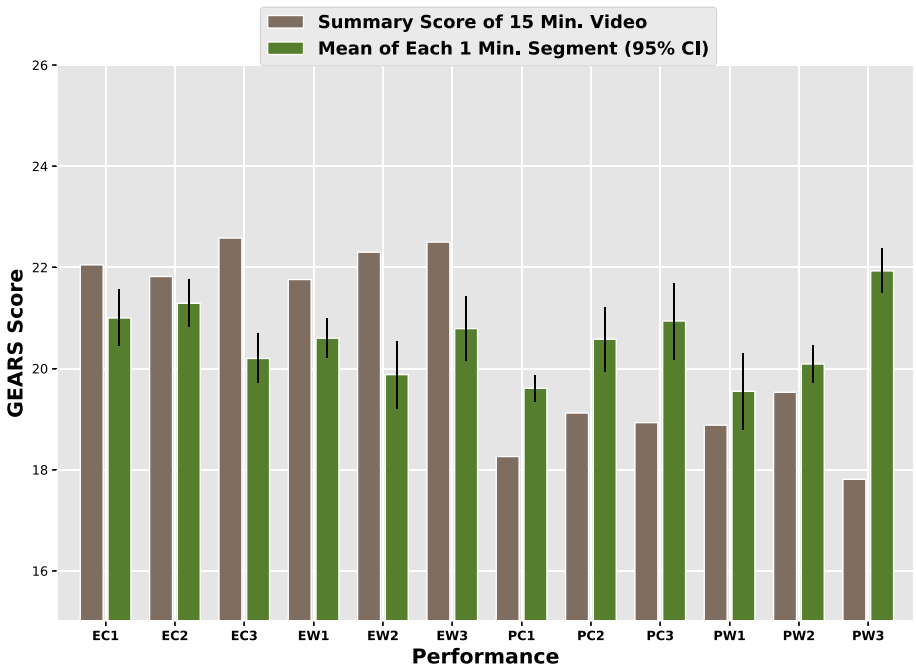

Fig. 3 Comparison of previously obtained scores for entire 15-min video to the mean score for all 15 1-min segments (95% CI) in the same video. The confidence intervals were not available from the vendor used to obtain scores for 15-min segments. (PW = proficient/warm-up; PC = proficient/control; EW = expert/warm-up; EC = expert/control)

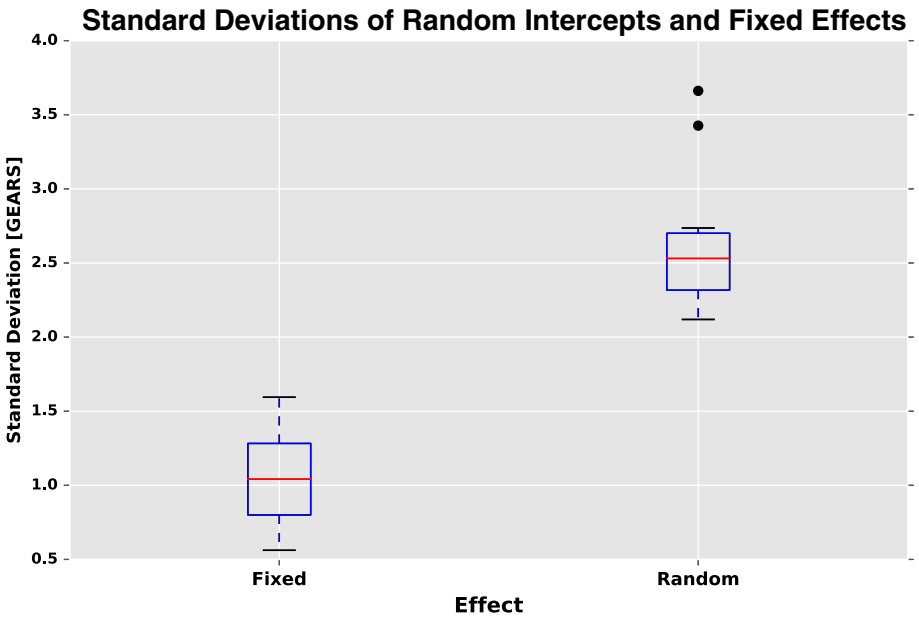

Fig. 4 Standard deviations of random intercepts and fixed-effect coefficients

vention, the difference in scores among the different time points was not significant [$\chi^2(14) = 1.856$, $p = 0.6179$].

A bar plot comparing the previously obtained summary score for each surgeon, grouped with that surgeon's weighted mean GEARS score averaged over all video segments and raters, is displayed in Fig. 3. Several of the previously obtained summary scores are outside the confidence interval band from the weighted mean GEARS scores for each video. The standard deviation of these weighted mean GEARS scores was 0.71 (95% CI from 0.50 to 1.20) (Fig. 4).

Conclusion

The results from the video segmentation study support our initial hypothesis that perceived surgeon skill may fluctuate throughout surgical procedures. If we consider the limitations of human attention span and time-varying fluctuations in skill for a 15-min video as contributing factors, it appears that human attention span limitations have a larger effect than natural variation of 'true' technical skill in time (i.e. measurement noise may be 2–3x larger than process noise). We found no compelling evidence that the underlying human skill (process) is constant for a 15-min duration in typical surgery. Perhaps crowds may be biased into giving scores that more closely reflect events they remember as being particularly good or bad. Given that scores varied from 2.11 to as many as 5.59 GEARS points throughout a video, this would clearly be a clinically relevant difference. A change in only a few points could easily be the difference between somebody being assessed as a novice instead of an expert surgeon.

This research may give credit to alternative surgical skill reporting methods, as opposed to giving static scores and labels. One alternative to this could be providing the mean of shorter segments of videos, as well as the amount of deviation in score throughout the performance, as a confidence interval. This would convey the typical skill of a surgeon in addition to how consistent their skill is through time.

This study includes a few limitations. There was no semantic segmentation of these tasks, only temporal 'chunking' into 1-min segments. Although videos of real surgeries were used, the data used were fairly sparse, with only 12 videos of 15-min surgeries. It is also not clear if the 1-min duration used for segmentation is optimal. Shorter or longer segments of surgery may be more beneficial to accurate skill evaluation. In this work, the 1-min ratings were obtained 3 months after the full-duration videos were rated. This may raise concerns about reliability (and comparability of data) of crowd ratings from studies separated by large time intervals.

Using these alternative reporting methods could aid supervised machine learning models in the future by reducing potential label noise. Static scores may lead to noisily labeled datasets, producing poorly trained classification models. Further study should consider optimal duration for human raters and a more rigorous analysis of the natural fluctuations of human technical skill.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical standard All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Funding This work was supported, in part, by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-15-2-0030, the National Science Foundation CAREER Grant under Award No. 1847610. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense, the National Science Foundation or the National Institutes of Health's National Center for Advancing Translational Sciences.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Kowalewski T, Comstock B, Sweet R, Schaffhausen C, Menhadji A, Averch T, Box G, Brand T, Ferrandino M, Kaouk J, Knudsen B, Landman J, Lee B, Schwartz BF, McDougall E, Lendvay TS (2015) Crowd-sourced assessment of technical skills for validation of basic laparoscopic urologic skills (BLUS) tasks. J Urol 195(6):1859–1865
2. Hung A, Chen J, Che Z, Nilanon T, Jarc A, Titus M, Oh PJ, Gill IS, Liu Y (2018) Utilizing machine learning and automated performance metrics to evaluate robot-assisted radical prostatectomy performance and predict outcomes. J Endourol 32(5):438–444
3. Makary M, Daniel M (2016) Medical error—the third leading cause of death in the US. BMJ 353:i2139
4. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, Birkmeyer NJ (2013) Surgical skill and complication rates after bariatric surgery. N Engl J Med 369(15):1434–1442
5. Kelly JD, Petersen A, Lendvay TS, Kowalewski TM (2020) The effect of video playback speed on surgeon technical skill perception. Int J Comput Assist Radiol Surg 15:739–747
6. Chen C, White L, Kowalewski T, Aggarwal R, Lintott C, Comstock B, Kuksenok K, Aragon C, Holst D, Lendvay T (2013) Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. J Surg Res 187(1):65–71
7. Doh FC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ (2012) Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. J Urol 187(1):247–252
8. Geri N, Winer A, Zaks B (2017) Challenging the six-minute myth of online video lectures: can interactivity expand the attention of learners? Online J Appl Knowl Manag 5(1):101–111
9. Schmitt M, Bulterman DCA, Cesar PS (2018) The contrast effect: QoE of mixed-video qualities at the same time. Qual User Exp 3(1):7
10. French A, Lendvay TS, Sweet RM, Kowalewski TM (2017) Predicting surgical skill from the first n seconds of a task: value over task time using the isogony principle. Int J Comput Assist Radiol Surg 12(7):1161–1170
11. Kelly JD, Heller N, Petersen A, Lendvay TS, and Kowalewski TM (2020) The effect of video playback speed on perception of technical skill in robotic surgery. Int J Comput Assist Radiol Surg (submitted)
12. Dockter R, Lendvay TS, Sweet RM, Kowalewski TM (2017) The minimally acceptable classification criterion for surgical skill: intent vectors and separability of raw motion data. Int J Comput Assist Radiol Surg 12:1151–1159
13. FFmpeg Developers (2016) ffmpeg tool (Version 4.1.3) [Software]. http://ffmpeg.org/
14. Core Team R (2018) R: a language and environment for statistical computing. Austria, Vienna. http://www.R-project.org/
15. Python Software Foundation. Python language reference, version 3.6. http://www.python.org