Toward training surgeons with motion-based feedback: Initial validation of smoothness as a measure of motor learning

Shivam Pandey[1], Michael D. Byrne[1,2], William H. Jantscher[3], Marcia K. O'Malley[2,3], Priyanshu Agarwal[3]
[1]Department of Psychology, [2]Department of Computer Science, [3]Department of Mechanical Engineering
Rice University, Houston, TX U.S.A.
{shivam.pandey, byrne, willam.jantscher, omalleym, pa20}@rice.edu

Surgery is a challenging domain for motor skill acquisition. A critical contributing factor in this difficulty is that feedback is often delayed from performance and qualitative in nature. Collection of high-density motion information may offer a solution. Metrics derived from this motion capture, in particular indices of movement smoothness, have been shown to correlate with task outcomes in multiple domains, including endovascular surgery. The open question is whether providing feedback based on these metrics can be used to accelerate learning. In pursuit of that goal, we examined the relationship between a motion metric that is computationally simple to compute—spectral arc length—and performance on a simple but challenging motor task, mirror tracing. We were able to replicate previous results showing that movement smoothness measures are linked to overall performance, and now have performance thresholds to use in subsequent work on using these metrics for training.

INTRODUCTION

Background

Training in many motor domains can be a challenge. Consider learning to do some motor task with a well-defined "success" vs. "failure" outcome metric, such as shooting a free throw in basketball. A trainee can shoot repeatedly and easily determine for each shot whether the shot was successful, but it may be very difficult for the trainee, and even a coach, to determine why different shots resulted in success or failure. In a domain like surgery, this is even more of a challenge, as "success" vs. "failure" may not be known for weeks or even months after a procedure is performed. Furthermore, the "coach" in surgery is typically one or more other surgeons, whose expertise is valuable and time spent training is time taken away from other activities.

Surgical skill is notoriously difficult to assess and evaluate (Moorthy, Munz, Sarker, & Darzi, 2003). The development of effective metrics to evaluate surgical skill is an active area of research (Reiley, Lin, Yuh, & Hager, 2011). The need for objective and quantitative assessment tools has been a topic of considerable interest and importance (Lin, Shafran, Yuh, & Hager, 2006; van Hove, Tuijthof, Verdaasdonk, Stassen, & Dankelman, 2010; Tsue, Dugan, & Burkey, 2007). Such need is driven by evidence that skill level can affect clinical outcomes after surgery (Reznick & MacRae, 2006). Assessment is often done informally through subjective feedback from other surgeons (Chaer et al., 2006; Bech et al., 2011; Riga et al., 2011) or based on a simple count of the number of times a procedure has been performed (Cronenwett, 2006; Schanzer et al., 2009).

Collection of high-density motion information, if the relevant metrics can be linked successfully with task outcomes, offers the opportunity to provide more detailed feedback that may speed the learning process. A quantitative and motion-based approach to performance assessment in manual control tasks is starting to gain traction in the research community, especially in the domain of robotic surgery and the corresponding simulation environments that are used to train surgeons in the use of the robotic technology. Specifically, access to higher quantities of more detailed data about the human's control over the task and the task outcomes provides the possibility to identify motion-based performance metrics that offer multiple advantages: insight into task performance, the ability to compare in a detailed manner the performance of trainees, and a mechanism to objectively track changes in performance as a result of training (i.e., learning curves).

In our prior work (Huegel, Celik, Israr, & O'Malley, 2009), we identified motion-based performance metrics that were associated with successful movement strategies for a virtual reality target-hitting task. After demonstrating that movement characteristics correlate with successful performance of the target- hitting task in terms of outcome measures, we were motivated to carry out these analyses for another more complicated motor control task. We developed a novel rhythmic motor control task in a simulated 3D virtual environment that is significantly more difficult to master than the target-hitting task (Howie, Purkayastha, Byrne, & O'Malley, 2011). This environment utilizes a high fidelity physics engine for rendering that task, and is unconstrained in that multiple movement strategies can result in successful completion of the task. We showed that motion-based performance metrics quantifying smoothness and user input frequency correlated with expert performance for this complex and unconstrained task (O'Malley, Purkayastha, Howie, & Byrne, 2014), using techniques similar to those in our prior work.

In our recent work (Estrada, O'Malley, Duran, Schulz, & Bismuth, 2014; Estrada, Duran, Schulz, Bismuth, Byrne, & O'Malley, 2016), we explored the applicability of motion-based measures of performance to endovascular surgery procedures. We evaluated performance in surgically relevant tasks specific to the endovascular domain in order to capture the unique characteristics of this specialty. We isolated specific tasks so as to reduce the risk of confounding our findings with assessment of procedural knowledge, and we evaluated performance among domain experts. The metrics were compared to structured grading assessments to determine

which metrics best captured skill. Our analysis showed that motion metrics based on principles of motor control that quantified smoothness (and in turn, quality of movement) were strongly correlated with the structured grading assessment, and were in fact superior to subjective structured grading assessments at discriminating expert from novice surgeons.

Motivation for Current Study

The fact that low-level motion metrics correlate with task outcomes is an important advance, but it raises additional questions, from theoretical questions about the nature of the relationship between those metrics and high-level outcomes to applied questions about the design of training systems. The ultimate goal of this research is to improve surgical task performance and the efficiency of training through motion-based feedback. The knowledge that motion-based metrics correlate with surgical performance is an important start, but it is only a starting point.

While we have established that motion-based metrics predict performance in multiple tasks, whether these metrics can be useful in training has not yet been established. If that can be established, then it raises the question of whether that feedback can be useful in surgical training in particular. Thus, there is still a long path between the knowledge that motion-based metrics are related to performance and using them as the basis for an intervention in surgical training.

The research presented here is the first step along that path. Before testing the use of feedback based on motion metrics on surgeons, we want to first establish that such metrics can be used to improve learning in a less specialized population with a less specialized task. This also raises questions about exactly how to use motion-based metrics to provide feedback, and what form that feedback should be and when it should be delivered.

The task: We wanted a task that stressed perceptual-motor performance like surgery, but one that has lower risk and can be attempted by anyone. The task needs to be challenging enough that there is ample room for improvement but not so daunting that subjects immediately give up. It would also be useful if this task is one where there is prior research that documented the learning curve so we have some idea what to expect. A task that meets all these criteria is mirror tracing (Snoddy, 1926). Our implementation of this task will be described in detail in the Method section.

Motion metrics: Our prior research showed that motion-based metrics based on principles of motor control that quantified smoothness—and in turn, quality of movement—were strongly correlated with the structured grading assessment. Submovement based metrics (number and duration of submovements) and spectral arc length (which evaluates smoothness in the frequency domain) showed the strongest and most significant correlations. Unfortunately, number of submovements is computationally extremely expensive to compute, and may not be suited for real-time feedback. Thus, in this experiment we examined only spectral arc length.

What is spectral arc length? Intuitively, consider what a smooth movement should entail: primarily low-frequency components. Conversely, a jerky movement will have larger amounts of high-frequency components. Spectral arc length is a method for quantifying this idea by looking at the complexity of the Fourier magnitude spectrum for the velocity profile of the movement. See (Balasubramanian, Melendez-Calderon, & Burdet, 2012) for details, including equations.

The first step to enabling feedback based on spectral arc length is that we need some basic data relating global performance on our mirror tracing task to the motion metrics computed for that task. That is, if we want to give subjects feedback on how well they performed the last trial, or last portion of a trial, based on the smoothness of movement, we need to know "how smooth is smooth?" for this task. More specifically, what specific values for spectral arc length should we use to trigger positive or negative feedback to subjects as they train? Before we can deliver feedback, we need ranges for these metrics that correspond to good and bad performance on the task.

Thus, the purpose of this experiment was simply to establish (1) that spectral arc length does correlate with performance on our mirror tracing task, and (2) what ranges of values are observed such that we can set bounds for feedback to be given in a subsequent training intervention.

METHOD

Subjects

There were 5 Rice undergraduate subjects, 1 male and 4 female, age range 18 to 20 years. Subjects were recruited from the Psychology subject pool and received credit toward a course requirement for participation.

Task

Subjects performed a computer version of the classic mirror tracing task pioneered by Snoddy (1926). In the original task, subjects used a metal stylus to trace around the interior of a physical six-pointed star made of brass, but subjects could not directly see either their hands or the star. Instead, they looked through a mirror, which reversed the left-right directional relationships between what subjects saw and how they actually moved, i.e., moving the stylus physically left appeared to move the stylus to the right.

Our version did not use a mirror, but rather presented the star on a computer display, as shown in Figure 1. The task was like the original, but rather than a stylus, subjects navigated a cursor around the star, and were instructed to keep the cursor inside the inner and outer boundaries. We paralleled the mirror by reversing the controller, but this time on both axes, so an upward movement of the controller moved the cursor down. Left and right directions were similarly reversed.

When subjects moved outside the boundary of the star the cursor changed color from green to red and all time spent outside the star incurred penalty time.

Design

Subjects performed five blocks of ten trials, so block and trial within a block were independent variables. There were no other variables manipulated. There were two dependent variables: time to complete a trial (including penalty time), and spectral arc length.

Procedure

Informed consent was obtained prior to the task. Each subject then performed 5 blocks of 10 trials each of the mirror tracing task, then completed a brief demographic survey and were debriefed.

On each trial, the cursor started in the circle at the left-hand vertex of the star, as depicted in Figure 1. The circle changed color from red to yellow to green, with green indicating that the subject should start moving clockwise around the star. As noted previously, when subjects moved outside the boundary of the star the cursor changed color from green to red and all time spent outside the star incurred penalty time. Penalty time was that time spent outside the star was charged at 3x time. That is, if they spent 3 seconds outside the star, 6 penalty seconds was added to their total completion time for the trial, meaning 9 total seconds were counted.

After each block of 10 trials, subjects were encouraged to take a short break before proceeding to the next block of trials.

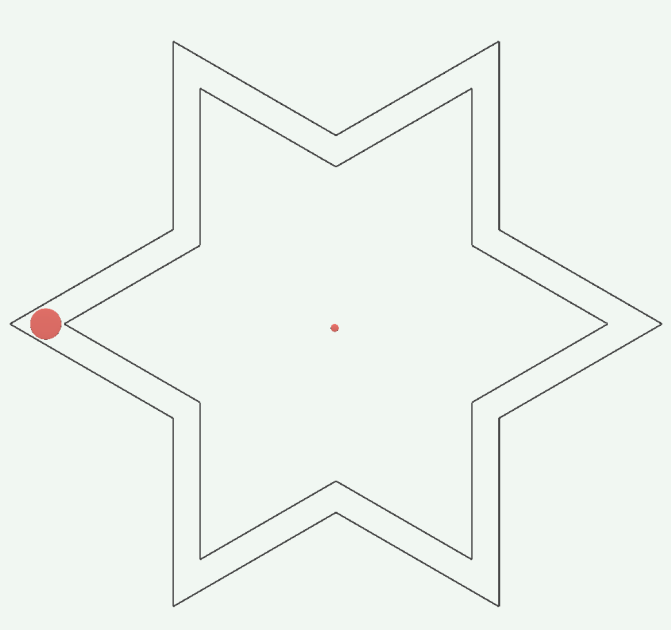

Figure 1. Display for the mirror tracing task

Materials

A Dell Optiplex 760 running Windows 7 and MATLAB 2015b was used to present the experiment on a Dell P2217 LCD monitor (55.87 cm or 22" diagonal) set to display at a resolution of 1680 by 1050 pixels.

Instead of a traditional mouse or joystick, subjects provided input via a Novint Falcon three-dimensional controller. The Falcon was set to input at only two dimensions that had a 10 cm x 10 cm physical workspace. Force feedback of approximately 9 Newtons was used to restrict movement to the two dimensions.

RESULTS

We collected a total of 250 trials of data (5 subjects, 50 trials each). Results for one trial were lost due to an unknown computer error. For the repeated-measures ANOVAs, the datum for this trial was replaced with the subject's mean.

The first question was whether the subjects showed the expected learning in overall task performance. They did; these data are presented in Figure 2. Total time on each trial was analyzed with a 5 (block) x 10 (trial within block) repeated-measures ANOVA, which showed only a reliable main effect of block $F(4, 16) = 8.39$, $MSE = 865$, $p = .024$ (Huynh-Feldt adjusted), Cohen's $f = 1.45$.

The second—and more interesting—question is the degree to which the spectral arc length measure was associated with improved overall task performance. The overall pattern was indeed quite similar; these data are presented in Figure 3. Spectral arc length was analyzed with the same 5 x 10 repeated-measures ANOVA which again showed only a main effect of block, $F(4, 16) = 7.35$, $MSE = 14.4$, $p = .019$ (Huynh-Feldt adjusted), Cohen's $f = 1.36$.

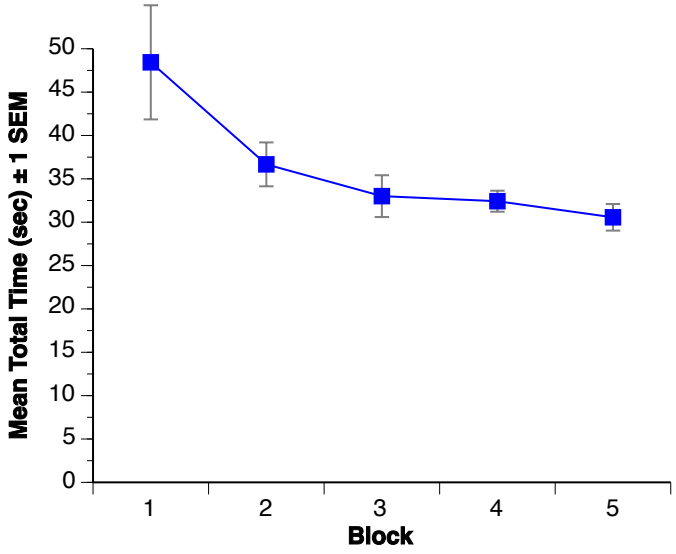

Figure 2. Overall task performance, as measured by total task time, as a function of block.

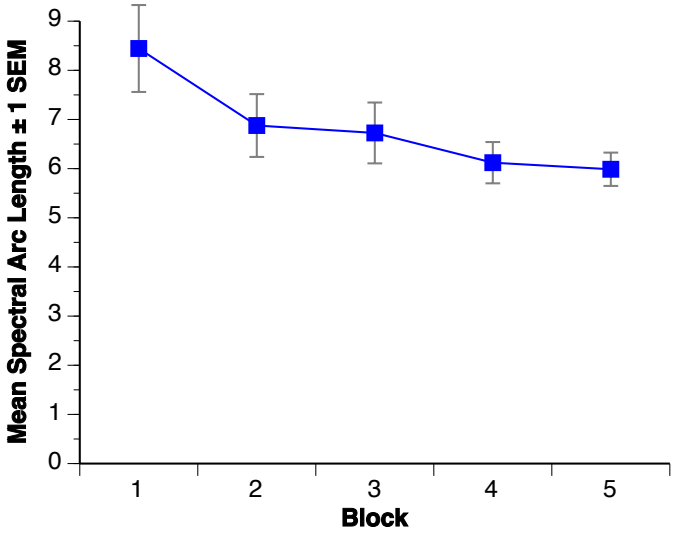

Figure 3. Spectral arc length as a function of block. Smaller values represent smoother movements.

These analyses suggest that, at least in relatively coarse terms, spectral arc length does indeed show the same pattern as overall task performance. However, more fine-grained analyses are appropriate here. This takes the form of looking at the relationship between task performance and spectral arc length on a per-trial basis. A scatter plot of all 249 valid data points is presented in Figure 4. As is apparent from the figure, as spectral arc length indicates smoother movement, overall task performance improves (that is, total time in the trial goes down). Overall, the association is fairly strong, $r = .75$ (significance testing here is not appropriate as these points are not all independent; this is presented simply as a descriptive measure).

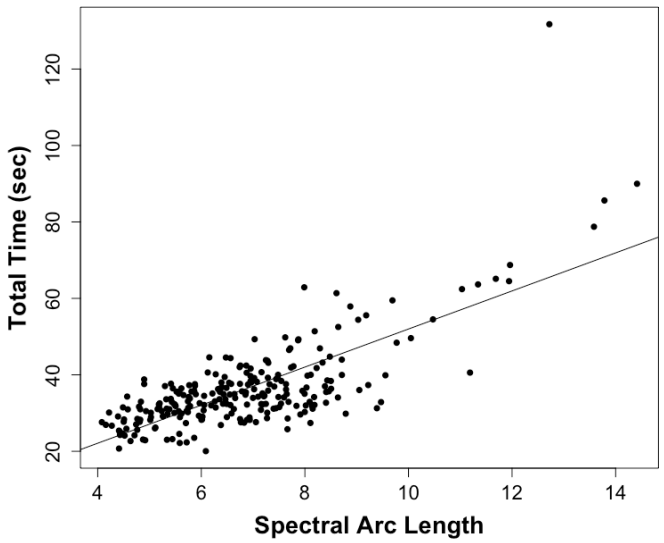

Figure 4. Total time in seconds vs. spectral arc length for all individual trials

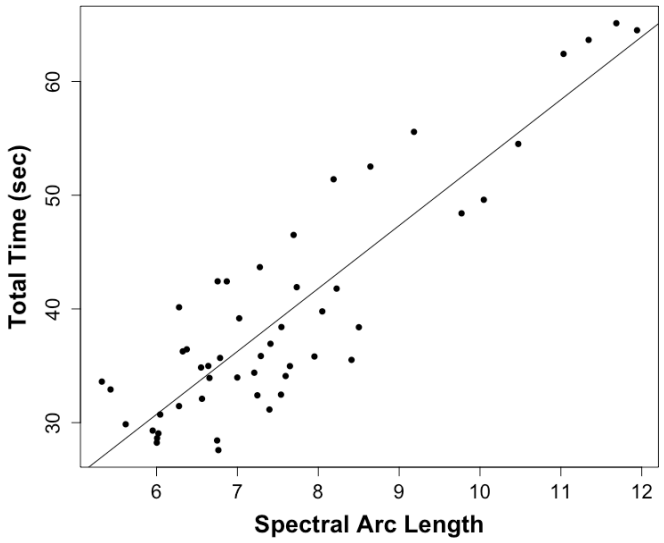

Figure 5. Total time in seconds vs. spectral arc length for one subject, strongest relationship

We also examined this relationship for each individual subject. The strongest relationship is shown in Figure 5, which shows $r = .89$. This subject never achieved particularly fast performance overall, and had few trials with an arc length of less than 6, but the coupling of arc length and overall performance was strong.

On the other hand, the subject with the weakest relationship is presented in Figure 6. This subject was much better at the task with about two-thirds of his/her trials completed in under 30 seconds and almost all arc lengths under 6. Furthermore, the relationship is not particularly strong, $r = .40$.

One possible explanation for the weak relationship for this subject is a range restriction effect. This particular subject overall showed little improvement in the task across the 50 trials, but his/her performance was, relative to the other subjects, consistently good—even on the first few trials. (This subject's slowest trial was still under 40 seconds.) Correspondingly, all spectral arc lengths were fairly low. In the larger picture, if the goal is to train subjects to produce good performance, this subject would require little training, and perhaps little feedback would be necessary to improve his/her performance.

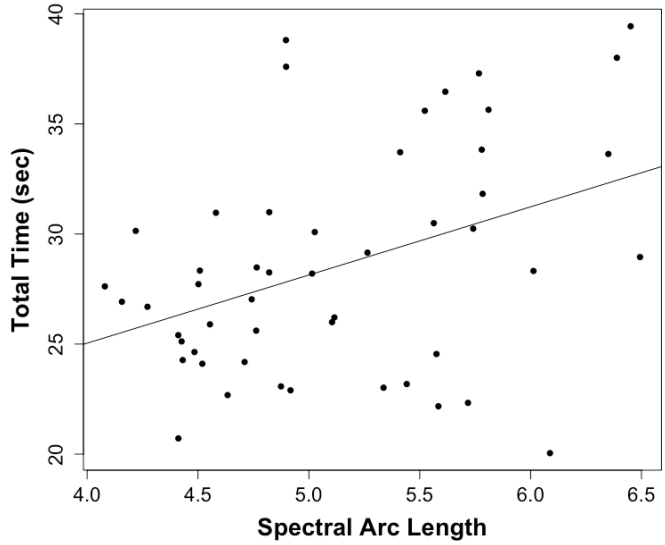

Figure 6. Total time in seconds vs. spectral arc length for one subject, weakest relationship

Overall, the fastest third of the trials performed roughly correspond to a spectral arc length of 6 or less, and arc lengths of 6–8 indicated middling performance, and spectral arc lengths of greater than 8 indicate poor performance. This suggests that the criteria for feedback should be at arc lengths of 6 and 8.

DISCUSSION

We had previously demonstrated a link between overall task performance and metrics derived from motion capture in several other domains, including endovascular surgery (Huegel et al., 2009; O'Malley et al., 2014; Estrada et al., 2016). Previous efforts, however, relied on using multiple motion metrics and seeing whether any of them showed such a relationship. This time, based on prior results and technical considerations, we selected a single metric. Fortunately, we

were able to replicate our previous findings; the motion metric showed a clear relationship with overall task performance.

The next step is to determine whether feedback based on this metric accelerates learning. All but one of our subjects showed substantial performance improvement over the course of the experiment (and the one subject who did not improve had rapid performance throughout). Thus, our plan going forward is to compute spectral arc length in real time and provide vibrotactile feedback either mid-trial (e.g., every 5 seconds or at every vertex of the star) or end-of-trial. Furthermore, we now have an idea of what values for spectral arc length should generate feedback: arc lengths of 8 or more should result in the sharpest feedback and arc lengths from 6 to 8 should produce intermediate feedback. Whether arc lengths of less than 6 should produce no feedback or positive feedback is an open question, as is the exact nature of the feedback itself. Vibrations can be varied in amplitude, frequency, and duration and so we will need to consider these properties as we move forward with the intervention. Of course, there are other critical issues that will also need to be investigated.

Feedback timing: In previous research, motion metrics were computed after subjects performed the task and then later compared to other performance metrics and used to predict the level of expertise of subjects. Doing the computation off-line and post task completion allows the use of computationally-intensive algorithms over large datasets. However, for training purposes, this is inadequate. Delaying feedback can decouple the feedback and the performance in ways that make it difficult to learn (Wickens, Hollands, Banbury, & Parasuraman, 2013). Thus, we want to deliver feedback close to real time, either mid-trial or end-of-trial.

Feedback delivery: Because surgery is a visually-intensive task and operating rooms are often loud environments, we decided to avoid visual and auditory channels for feedback delivery and will deliver performance feedback haptically. Our plan is to use a C2 tactile feedback device (Engineering Acoustics, Inc.) worn above the elbow in a portable music player armband. This will allow us to deliver feedback in a modality that should not interfere with other modalities critical to surgery. (Note that feedback would be delivered to the non-dominant arm in order to not interfere with the actual surgical motions.)

Ultimately, if we can show that the intervention can indeed accelerate learning, then we will move away from undergraduate subjects and mirror tracing to see whether the same kinds of interventions can also speed up the rate of acquisition of a much more complicated motor skill, endovascular surgery.

ACKNOWLEGEMENTS

This research was supported by grant #IIS-1638073 from the National Science Foundation. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies or endorsements, either expressed or implied, of NSF, the U.S. Government, or any other organization.

REFERENCES

Balasubramanian, S., Melendez-Calderon, A., & Burdet, E. (2012). A robust and sensitive metric for quantifying movement smoothness. *IEEE Transactions on Biomedical Engineering, 59(8)*, 2126–2136.

Bech, B., Lönn, L., Falkenberg, M., Bartholdy, N.J., Rader, S.B., Schroeder, T.V., & Ringsted, C. (2011). Construct validity and reliability of structured assessment of endovascular expertise in a simulated setting." *European Journal of Vascuclar and Endovascular Surgery, 42(4)*, 539-548.

Chaer, R. A., Derubertis, B. G., Lin, S. C., Bush, H.L., Karwowski, J. K., Birk, D., Morrissey, N.J., Faries, P. L., McKinsey, J. F., & Kent, K.C. (2006). Simulation improves resident performance in catheter-based intervention: Results of a randomized, controlled study. *Annals Surgery, 244(3)*, 343-352.

Cronenwett, J.L. (2006). Vascular surgery training: Is there enough case material? *Seminar Vascular Surgery, 19*, 187-90.

Estrada, S., Duran, C., Schulz, D., Bismuth, J., Byrne, M. D., & O'Malley, M. K. (2016). Smoothness of surgical tool tip motion correlates to skill in endovascular tasks. *IEEE Transactions on Human-Machine Systems*, *46*, 647–659.

Estrada, S., O'Malley, M.K., Duran, C.A., Schulz, D.G. & Bismuth, J. (2014). On the development of objective metrics for surgical skills evaluation based on tool motion. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, San Diego, CA, October 5-8.

Huegel, J.C., Celik, O., Israr, A., & O'Malley, M.K. (2009). Expertise-based performance measures in a virtual training environment. *Presence, 18(6)*, 449–467.

Howie, N., Purkayastha, S. N., Byrne, M. D., & O'Malley, M. K. (2011). Motor skill acquisition in a virtual gaming environment. In *Proceedings of the Human Factors and Ergonomics Society 55th Annual Meeting,* (pp. 2148–2152). Santa Monica, CA: Human Factors and Ergonomics Society.

Lin, H.C., Shafran, I., Yuh, D., & Hager, G. D. (2006). Towards automatic skill evaluation: detection and segmentation of robot-assisted surgical motions. *Computer Aided Surgery, 11(5)*, 220–230.

Moorthy, K., Munz, Y., Sarker, S.K. & Darzi, A. (2003) Objective assessment of technical skills in surgery. *British Medical Journal, 327*, 1032–1037.

O'Malley, M. K., Purkayastha, N., Howie, N., & Byrne, M. D. (2014). Identifying successful motor task completion via motion-based performance metrics. *IEEE Transactions on Human-Machine Systems, 44*, 139–145.

Reiley, C., Lin, H.C., Yuh, D.D., & Hager, G.D. (2011). Review of methods for objective surgical skill evaluation. *Surgical Endoscopy, 25(2)*, 356–366.

Reznick, R.K. & MacRae, H. (2006). Teaching surgical skills? Changes in the wind. *New England Journal of Medicine, 355(25)*, 2664–2669.

Riga C.V., Bicknell, C.D., Hamady, M.S., & Cheshire, N.J.W. (2011) Evaluation of robotic endovascular catheters for arch vessel cannulation. *Journal of Vascular Surgery Official Publication Society Vasculary Surgery & International Society Cardiovascular Surgery North American Chapter, 54(3)*, 799–809.

Schanzer, A., Steppacher, R., Eslami, M., Arous, E., Messina, L., & Belkin, M. (2009) Vascular surgery training trends from 2001-2007: A substantial increase in total procedure volume is driven by escalating endovascular procedure volume and stable open procedure volume. *Journal of Vasclar Surgery, 49*, 1339–1344.

Snoddy, G. S. (1926). A psychophysiological analysis of a case of motor learning with clinical applications. *Journal of Applied Psychology, 10(1)*, 1-36.

Tsue, T.T., Dugan, J. W., & Burkey, B. (2007). Assessment of surgical competency. *Otolaryngologic Clinics of North America, 40(6)*, 1237–1259.

Van Hove, G.J.M., Tuijthof, P. D., Verdaasdonk, E.G.G., Stassen, L. P. S., & Dankelman, J. (2010). Objective assessment of technical surgical skills. *British Journal of Surgery, 97(7)*, 972–987.

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2013). *Engineering psychology and human performance.* Boston: Pearson.