Developing Effective Automated Feedback for Temporal Bone Surgery Simulation

Sudanthi Wijewickrema[1](BEng(Hons), PhD), Patorn Piromchai[1](MD, MSc),

Yun Zhou[2](BSc, MSc), Ioanna Ioannou[1](BEng/BCom), James Bailey[2](BEng, PhD),

Gregor Kennedy[3](BE, PhD), Stephen O'Leary[1](MBBS, FRACS, PhD)

1. Department of Otolaryngology, University of Melbourne, Australia

2. Department of Computing & Information Systems, University of Melbourne, Australia

3. Centre for the Study of Higher Education, University of Melbourne, Australia

Corresponding author:

Sudanthi Wijewickrema

Dept. of Otolaryngology, University of Melbourne,

Level 2, Royal Victorian Eye and Ear Hospital,

32, Gisborne Street, East Melbourne, VIC 3002,

Australia

Tel - +61 3 9929 8386

Email – swijewickrem@unimelb.edu.au

Structured Abstract

Objective: We aim to facilitate effective surgical skill learning in virtual reality simulation-based training environments using automated real-time feedback.

Study Design: We introduce a feedback system that emulates the advice of a human expert based on a multivariate analysis of drilling behaviour within a temporal bone surgery simulator. We evaluate its performance through a study of 24 medical students (12 with feedback and 12 without) performing virtual cortical mastoidectomy.

Setting: The feedback system was based on the Melbourne University virtual reality temporal bone surgery simulator. The study was performed at the simulation laboratory of the Royal Victorian Eye and Ear Hospital, Melbourne.

Subjects and Methods: The study participants were medical students from the University of Melbourne. The extent to which the drilling behaviour of the feedback and non-feedback groups differed was used to evaluate the effectiveness of the system. Its accuracy was determined through a post-experiment observational assessment of recordings made during the experiment by an expert surgeon. Its usability was evaluated using students' self reports of their impressions of the system.

Results: A Friedman's test showed that there was a significant improvement in the drilling performance of the feedback group. The post-experiment assessment demonstrated that the system provided timely feedback 88.6% of the time and appropriate feedback 84.2% of the time. Participants' opinions about the usefulness of the system were highly positive.

Conclusion: The automated feedback system was observed to be effective in improving surgical technique and the provided feedback was found to be accurate and useful.

Introduction

Apprenticeship has long been the backbone of surgical education, where an expert provides the trainee with feedback during supervised operative activity. More recently, competence-based training has played a significant role in response to calls for structured learning, reduced opportunities for operating room training and community insistence on greater accountability of training programs. Within this context, simulation has emerged as an important training tool[1,2]. Virtual reality (VR) training environments are seen as advantageous because they allow repeated training in risk-free environments. They are particularly useful in domains such as surgery, where training resources are limited, participant numbers are high, and failure is either expensive or catastrophic. Similar approaches have been used in training for aviation[3], health[1], defense[4], and emergency services[5].

Simulation can be used to support the educational principle of "deliberate practice"; the concept that in order for a novice to become an expert, he/she is required to undertake tasks with the explicit intent of improving his/her skills[6]. Deliberate practice calls for the individual to focus on a defined task, typically identified by a teacher, to improve particular aspects of performance; it involves repeated practice along with coaching and immediate feedback on performance[7]. Typically, the onus of providing feedback falls on human experts, and the need for them to oversee the training greatly limits the utility and application of VR training environments.

Previous attempts at overcoming the need for expert supervision in VR training environments have mostly focused on end-of-task summative assessment[8-10]. While summative feedback may be constructive, it cannot replace meaningful feedback provided during training. The few researchers who have looked into the provision of real-time automated feedback in surgical

simulation have only provided relatively simple forms of feedback. For example, Rhienmora at al.[11] provided real-time feedback on individual metrics (force, position, and orientation) in a dental simulator by making comparisons with average expert values. Fried et al.[12] quantitatively defined a range of errors for surgical performance (violation of tissue, violation of instrument tolerances, force patterns, etc.) and provided real-time feedback by making comparisons with a database of metrics from pre-recorded performances in an endoscopic sinus surgery simulator. Sewell et al.[13] provided real-time feedback on bone removed with correct/incorrect technique according to the currently selected metric (visibility, force, or removal region) in the form of coloured voxels (3D points) in a temporal bone surgery simulator. Our previous work[14] provided automated feedback on force applied by trainees performing virtual temporal bone surgery.

Typically, these systems provided real-time feedback based on the analysis of individual metrics. However, surgical skill is multi-faceted and there exist complex interactions between various metrics that define it[15,16]. Moreover, feedback based on univariate analyses do not closely emulate the meaningful and nuanced advice that human experts provide during surgical training. We attempt to bridge this gap by introducing a system that provides real-time feedback on surgical technique based on multi-dimensional models of surgical expertise as applied to virtual temporal bone surgery. The system was trained to classify hand movements of surgeons as “expert” or “trainee” drilling behavior, and deliver feedback when trainee drilling was observed. The feedback consisted of advice on how to modify specific aspects of the drilling technique to better approximate expert behaviour and warnings when trainees approached a critical anatomical structure with the drill.

In the study reported here, medical students undertook cortical mastoidectomy, the

foundational operation on the temporal bone, within the virtual environment. After receiving standardized instructions on conducting the surgery, they were randomly allocated to receiving automated feedback or not. The main aim of the study was to determine whether participants receiving feedback significantly modified their drilling technique to approximate expert behaviour and avoid injury to critical anatomical structures, compared to those receiving no feedback. A secondary aim was to determine whether the feedback given was appropriate and timely. Finally, the study aimed to determine the usability and usefulness of the feedback as assessed by participants' self reports of their impressions of the system.

Methods

Test Platform

The simulation environment used in this research was the University of Melbourne's VR temporal bone surgery simulator[17]. With this simulator, surgeons can practice otological operations such as mastoidectomy, middle ear surgery and the approach to cochlear implantation. The simulator presents the trainee with two slightly offset images to produce the illusion of a 3D operating space, when viewed through 3D glasses (see Figure 1). Major anatomical structures that must be identified without injury during surgery, such as the facial nerve, sigmoid sinus, dura, ossicles and the labyrinth are represented in the virtual temporal bone. The surgeon interacts with the virtual temporal bone using a pen-like haptic device (surgical drill) that provides force feedback in three dimensions.

Design of the Feedback System

To provide surgical technique feedback, we trained a classifier to recognize expert and trainee behaviour using a previously collected dataset of 16 performances provided by 7 experts and 11 performances provided by 6 trainees on the simulator. The training data consisted of a

series of “strokes” identified in the continuous data stream output by the simulator during a surgical task. A stroke was defined as a set of points representing a continuous drilling motion. The end of a stroke was considered to be reached when there was no material being removed or when the direction of the trajectory showed an abrupt change[18]. For each stroke, metrics that represent surgical technique (duration, length, average speed, average acceleration, average force, straightness, median burr size, average magnification level, bone removal rate, and average distance to closest anatomical structure) were determined[19-21]. The strokes obtained from the expert and trainee performances were used to build a model that identifies expert and trainee behaviour[16].

The model, once trained could be presented with new data to be classified according to the patterns detected during training. If a stroke with poor surgical technique was detected, advice on how to improve the performance could be provided. To this end, the feedback system determined the best metric to provide advice on, such that surgical technique could approach the expert ideal. Thus, surgical technique feedback took the form of a suggestion, delivered via audio, to either increase or decrease a metric such as stroke length, stroke speed, stroke straightness, force, burr size, or magnification level[16].

Proximity feedback was provided when the drill tip came within 10mm of an anatomical structure. The aim of this feedback was to make the trainees aware that they were nearing a structure and remind them to exercise caution when drilling, so as to expose the structure without causing critical damage (e.g. facial paralysis, intracranial injury, severe hemorrhage, or deafness).

When providing feedback, repetitions and delays were used to ensure accuracy and to avoid

overloading the user with feedback. Surgical technique feedback was only provided to the user after detecting n repetitions of the same feedback. Once a feedback was presented to the user, processing of strokes was paused for a t period of time. If the same feedback was detected within a T time period after the previously presented feedback, it was ignored. In our trials, n=2, t=5s, and T=10s were established to be optimal values for the system. Figure 2 illustrates the workflow of the feedback system (see our previous work[22] for more details).

Experimental Setup

To evaluate the performance of the feedback system, 24 students were recruited (13 MBBS, 10 MD, and 1 PhD) to participate in an experimental study. This study protocol was approved by the Human Research Ethics Committee of the University of Melbourne (HREC #1135497). All participants had prior knowledge of the anatomy of the ear, but had no surgical experience. They were shown a video tutorial on how to perform a cortical mastoidectomy, taught how to use the simulator, and after a familiarization period, asked to perform this procedure on the simulator twice. 12 participants were provided with automated real-time feedback, while the remaining 12 participants were not provided with feedback in this form. The performance of all participants was recorded using a continuous data stream from the simulator and through the use of screen capture software. At the end of the procedure, participants in the feedback group were interviewed to obtain their views on the system.

To evaluate the effectiveness of the feedback in modifying stroke technique, the percentage of strokes classified as expert (using the behaviour model discussed above) for the two groups was compared using a Friedman's test. Further analyses of how the stroke technique changed between the groups at different stages of the surgical procedure were also conducted.

A post-experiment evaluation carried out by an expert otologist assessed the accuracy of the feedback system on three error measures: 1) “false positive” classifications: when feedback was provided while stroke technique was acceptable, 2) “wrong feedback”: when participants’ technique was accurately classified as “trainee” but the content of the feedback was inaccurate, and 3) “false negative” classifications: when feedback was not provided while stroke technique was unacceptable.

The metrics used to define surgical technique (stroke duration, stroke length, speed, acceleration, force, straightness of stroke, size of the burr, magnification level, bone removal rate, and distance to anatomical structures) were compared between groups using a Friedman’s test to assess how they were affected by the feedback.

The amount of damage caused to anatomical structures was compared for the two groups in an attempt to evaluate the effectiveness of the proximity feedback. The damage caused was measured as the percentage of structure voxels (i.e. voxels that make up critical anatomical structures) drilled when compared to the total number of voxels drilled during the procedure. Further, the end products of the procedures of all participants were evaluated by an expert otologist using the Welling Scale[23], a validated method of assessment of the quality of a mastoidectomy that systematically scores exposure and injury of key surgical landmarks. A Friedman’s test was performed on the resulting scores to identify differences in the performance of the two groups.

To assess the usability and usefulness of the feedback system, the participants’ answers to the following interview questions were analyzed: 1) did you pay attention to the feedback and

notice it while you completed the task?, 2) did it assist you when you were completing the procedure or stages of it?, 3) was it unhelpful, irrelevant or distracting when you were completing the procedure or stages of it?, and 4) how could the provision of feedback by the system be improved or be made more useful?

Results

The results of the Friedman's test, after adjusting for the effects of repetition, showed that the percentage of expert strokes of the feedback group was significantly higher than that of the non-feedback group ($\chi^2(1) = 14.450$, $p < 0.001$). A post-hoc analysis of the data using a Bonferroni adjustment showed that there was no significant difference between the two repetitions. Given the lack of difference in stroke technique between simulation procedures, the data for each participant across the two repetitions were combined (averaged). The percentage of expert strokes in the two groups during different stages (at 10% intervals of completion) showed a consistent difference throughout the procedure (see Figure 3).

A total of 576 feedback messages were provided across the two repetitions of the participants in the feedback group. 39 feedback messages were determined to be false positives; 52 messages were assessed as wrong feedback; and 69 instances were identified as false negatives where feedback should have been provided but wasn't. Therefore, timely feedback was provided by the system 88.6% of the time, and in 84.2% of these instances it was accurate.

Out of all the metrics used to define stroke technique, only the bone removal rates, when removing either solid cortical bone or porous trabeculated bone, were found to be significantly different between the two groups ($\chi^2(1) = 4.050$, $p = 0.044$ and $\chi^2(1) = 6.050$, $p =$

0.014 respectively) after adjusting for effects introduced by the repetitions. The bone removal rates were found to be higher in the feedback group when compared to the control group.

No significant differences were observed either between groups or between participants' repetitions with respect to the percentage of structure voxels damaged or the scores obtained using the Welling Scale[23].

The majority of the participants indicated that they noticed the feedback, paid attention to it when completing the task, and found it to be useful. Participants commented particularly on the helpfulness of the warnings provided when they were close to an anatomical structure. For example, participant P06 stated: "it reminded me to be gentle near structures". Feedback on surgical technique was also deemed to be helpful. For example, participant P01 said: "particularly helpful was changing burr size and whether or not to zoom in". Only one participant indicated that the feedback was unhelpful while a few mentioned it was distracting at times. For example, P01 said: "sometimes it was really out of the blue and caught you off guard".

Discussion

The results of the study indicate that the surgical technique feedback offered by the system was effective in guiding drilling behavior towards an "expert" ideal, demonstrating that it can help shape better surgical technique. These results are consistent with those of previous studies on automated feedback within surgical simulators[13,24], although none were conducted on feedback of the sort provided in this study. The major difference was that in our study, the focus was on surgical technique as a multivariate behavior model rather than on the individual metrics that define it. For example, Sewell et al.[13] provided real-time feedback on bone

removal (based on individual metrics such as visibility, force, and removal region) and showed that the group that received feedback maintained better visibility while drilling. Judkins et al.[24] presented feedback on parameters such as speed, grip force, and relative phase in virtual laparoscopy surgery and observed that it improved performance with respect to these metrics.

Whether the observed modification of drilling behaviour is in effect a move towards improved performance is largely dependent on how close the expert ideal is to actual expert performance, and if such an ideal can in fact be defined for a given task. It has been shown previously that expert and trainee behaviour can be differentiated on a virtual reality temporal bone simulator[19] and the core metrics that define this difference have also been identified[20,21]. The multivariate behaviour models developed using these metrics have been shown to accurately classify expert and trainee performance[16] indicating that for our application, an expert ideal can be defined. These results are further validated by the outcome of the post-experiment analysis that assessed the accuracy of the feedback system.

The need for using multivariate over univariate drilling behavior to define the ideal is that there are interactions between individual metrics that are not expressed in a univariate model (e.g. the relationship between force and proximity to structures identified in Wan et al.[25] as an element of competency evaluation in cortical mastoidectomy). There is a risk that if individual dimensions of expert drilling behavior were presented to trainee surgeons, they may concentrate upon these at the exclusion of others, running the risk that other aspects of surgical technique are compromised. For example, novices aspiring to emulate the faster drilling rates of experts may ignore other factors such as force, orientation of the drill, and proximity to structures, thereby risking injury to critical anatomy.

Of the core metrics used to define surgical technique (ie. stroke duration, stroke length, speed, acceleration, force, straightness of stroke, size of the burr, magnification level, bone removal rate, and distance to anatomical structures), only bone removal rate was significantly different between the groups. This demonstrates that greater efficiency in drilling was achieved by the feedback group without evidence of increased damage to structures or reduced scores in the end-product analysis. However, previous studies[19,26] show that other independent metrics such as force are also useful in differentiating the experience levels of surgeons. This implies that the changes in the surgical technique of participants who received feedback were not large enough to be detected by a univariate analysis of individual metrics. It is also probable that, as rank beginners with no training in operative surgery, the participants were not able to build enough competence to reach a higher skill level. This outcome is consistent with educational notions of deliberate practice, which indicates that the number of hours spent in practice is an important determinant of the level of expertise[27]. The observations that proximity warnings did not significantly reduce anatomical structure damage, and that the end product assessment scores were not significantly different between groups complement the above observations and also suggest that expertise has more dimensions that have not been addressed in this study. Indeed, a comprehensive surgical training program would require feedback to be provided in dimensions beyond psychomotor skills (which was the main focus of this study), such as advice on where to drill and how to proceed at certain points of the procedure.

According to the outcomes of this study, surgical drilling behaviour of medical students could be improved by providing automated real-time feedback. However, it was observed that more practice is required to effectively improve their overall performance. Given the encouraging

findings presented here, future studies are warranted to probe issues relating to skill retention, and effects of repeated practice over time. Studies in other fields (e.g. in laparoscopy[28,29]) suggest that skills learned through simulation-based training do successfully transfer to the operating room, and it would be of interest to determine whether the same is found when automated feedback is integrated into the simulation environment.

In conclusion, the results of this study suggest that trainees could be guided in the "right" direction when learning to handle a drill in a surgical procedure, and indicate that the dream of a self-guided simulation-based surgical training system for temporal bone surgery is attainable. Such a training platform would not only reduce the burden placed on expert instructors, but would also assist in producing a better class of surgeons.

Acknowledgements

The authors would like to thank the Asian Office of Aerospace Research and Development and Cochlear Ltd. for funding this project.

Figures

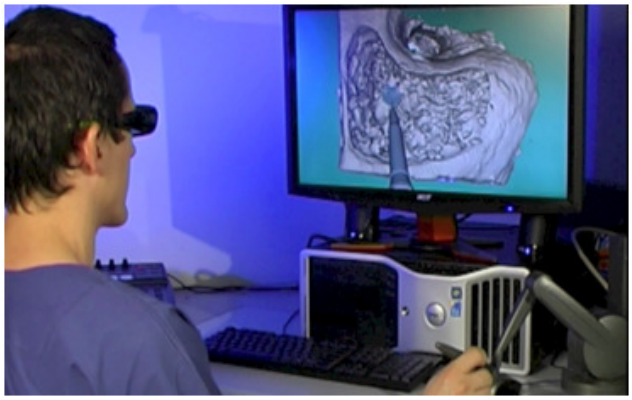

Figure 1: The Melbourne University Virtual Reality Temporal Bone Surgery Simulator

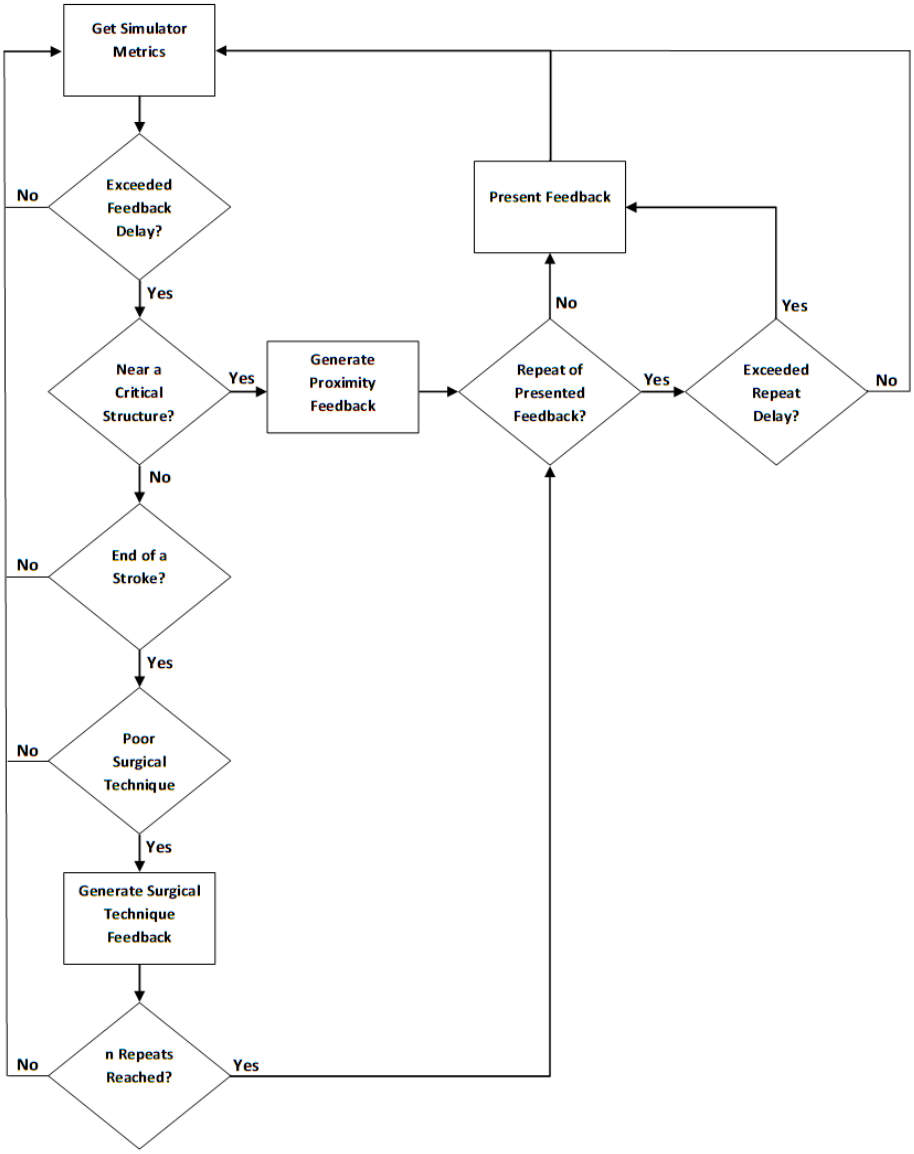

Figure 2: Design of the Feedback System

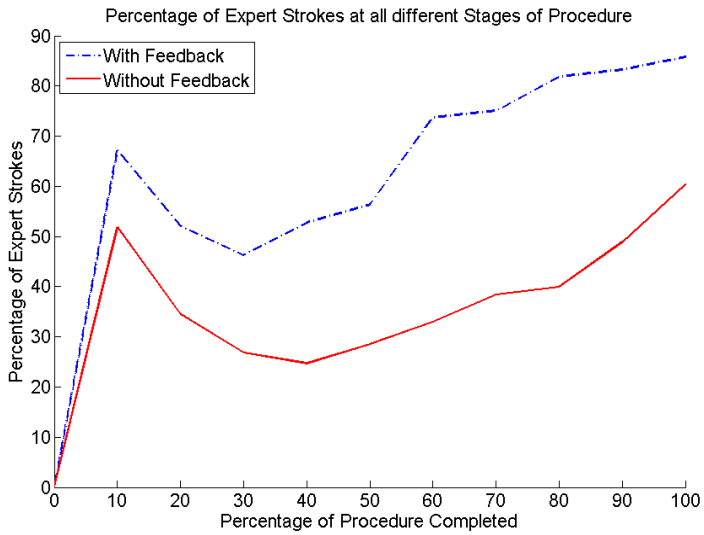

Figure 3: Comparison of Surgical Technique at Different Stages of the Procedure

References

1. Hammoud MM, Nuthalapaty FS, Goepfert AR, et al. To the point: medical education review of the role of simulators in surgical training. *American Journal of Obstetrics & Gynecology.* 2008;199(4):338-343.

2. Chen C-H, Jeng M-C, Fung C-P, et al. Psychological Benefits of Virtual Reality for Patients in Rehabilitation Therapy. *Journal of Sport Rehabilitation.* 2009;18(2):258-269.

3. Howard, CE. Simulation & training: expecting the unexpected. *Military & Aerospace Electronics*. 2011;22(11): 12-23.

4. Cosma, D, Stanic, M-P. Implementing a Software Modeling-Simulation in Military Training. *Revista Academiei Fortelor Terestre.* 2011;16(2): 204-215.

5. Syi, S, Shih, C-L. Modeling an emergency medical services system using computer simulation. *International Journal of Medical Informatics.* 2003;72(1-3): 57-72.

6. Ericsson KA, Krampe RT, Tesch-Römer C. The role of deliberate practice in the acquisition of expert performance. *Psychological Review.* 1993;100(3):363-406.

7. Cox M, Irby DM, Reznick RK, et al. Teaching surgical skills—changes in the wind. *New England Journal of Medicine.* 2006;355(25):2664-2669.

8. Macke T, Rosen J, Pugh C. Data Mining of the E-Pelvis Simulator Database: A Quest for a Generalized Algorithm for Objectively Assessing Medical Skill. *Medicine Meets Virtual Reality 14: Accelerating Change in Healthcare: Next Medical Toolkit* 2005:355-360.

9. Megali G, Sinigaglia S, Tonet O, et al. Modelling and Evaluation of Surgical Performance Using Hidden Markov Models. *Biomedical Engineering, IEEE Transactions on.* 2006;53(10):1911-1919.

10. Kerwin T, Wiet G, Stredney D, et al. Automatic scoring of virtual mastoidectomies

using expert examples. *Int J CARS.* 2012/01/01 2012;7(1):1-11.

11. Rhienmora P, Haddawy P, Suebnukarn S, et al. Intelligent dental training simulator with objective skill assessment and feedback. *Artificial Intelligence in Medicine.* 2011;52(2):115-121.

12. Fried MP, Satava R, Weghorst S, et al. Identifying and reducing errors with surgical simulation. *Quality and Safety in Health Care.* October 1, 2004 2004;13(suppl 1):i19-i26.

13. Sewell C, Morris D, Blevins NH, et al. Providing metrics and performance feedback in a surgical simulator. *Computer aided surgery: official journal of the International Society for Computer Aided Surgery.* 2008;13(2):63.

14. Kennedy G, Ioannou I, Zhou Y, et al. Mining interactions in immersive learning environments for real-time student feedback. *Australasian Journal of Educational Technology.* 2013;29(2).

15. Sewell C, Morris D, Blevins NH, et al. Validating metrics for a mastoidectomy simulator. *Stud Health Technol Inform.* 2007;125:421-426.

16. Zhou Y, Bailey J, Ioannou I, et al. Constructive Real Time Feedback for a Temporal Bone Simulator. *International Conference on Medical Image Computing and Computer Assisted Intervention.* 2013;315-322.

17. O'Leary SJ, Hutchins MA, Stevenson DR. , et al. Validation of a Networked Virtual Reality Simulation of Temporal Bone Surgery. *The Laryngoscope.* 2008;118(6):1040-1046.

18. Hall R, Rathod H, Maiorca M, et al. Towards Haptic Performance Analysis Using K-Metrics. *Haptic and Audio Interaction Design.* 2008;50-59.

19. Zhao YC, Kennedy G, Hall R, et al. Differentiating levels of surgical experience on a virtual reality temporal bone simulator. *Otolaryngology-Head and Neck Surgery.*

2010;143(5):S30-S35.

20. Zhao YC, Kennedy G, Hall R, et al. Which computer based metrics are most predictive of expertise in a virtual reality temporal bone simulator? *Conference of the Australian Society of Otolaryngology Head and Neck Surgery.* 2010.

21. Zhao YC, Kennedy G, Hall R, et al. What training benefit can haptic-enabled virtual reality simulation contribute towards surgical education in temporal bone surgery? Results from a randomised control trial. *The Conference of the Australian Society of Otolaryngology Head and Neck Surgery.* 2010.

22. Wijewickrema S, Ioannou I, Zhou Y, et al. A Virtual Reality Temporal Bone Surgical Simulator with Automated Real-Time Feedback for Effective Learning of Surgical Technique. *Medicine meets Virtual Reality.* 2014;In Press.

23. Butler NN, Wiet GJ. Reliability of the Welling scale (WS1) for rating temporal bone dissection performance. *Laryngoscope.* Oct 2007;117(10):1803-1808.

24. Judkins TN, Oleynikov D, Stergiou N. Enhanced robotic surgical training using augmented visual feedback. *Surgical innovation.* 2008;15(1):59-68.

25. Wan D, Wiet GJ, Welling DB, et al. Creating a cross-institutional grading scale for temporal bone dissection. *The Laryngoscope.* 2010;120(7):1422-1427.

26. Agus M, Giachetti A, Gobbetti E, et al. A Haptic Model of a Bone-cutting Burr. *Studies in Health Technology and Informatics.* 2003; 94:4-10.

27. Ericsson KA. The acquisition of expert performance. *The road to excellence.* 1996:1-50.

28. Korndorffer Jr., JR, Dunne JB, Sierra R, et al. Simulator training for laparoscopic suturing using performance goals translates to the operating room. *Journal of the American College of Surgeons.* 2005;201(1):23-29.

29. Torkington J, Smith S, Rees B, Darzi A. Skill transfer from virtual reality to a real laparoscopic task. *Surgical Endoscopy.* 2001;15(10):1076-1079.