The Journal of Laryngology & Otology

cambridge.org/jlo

Automated summative feedback improves performance and retention in simulation training of mastoidectomy: a randomised controlled trial

A Frithioff[1,2], M Frendø[1,2], J Hastrup von Buchwald[1,2], P Trier Mikkelsen[3], M Sølvsten Sørensen[1] and S Arild Wuyts Andersen[1,2]

[1]Department of Otorhinolaryngology – Head and Neck Surgery and Audiology, Rigshospitalet, Copenhagen, Denmark, [2]Copenhagen Academy for Medical Education and Simulation, The Capital Region of Denmark, Aarhus, Denmark and [3]The Alexandra Institute, Aarhus, Denmark

Main Article

Dr A Frithioff takes responsibility for the integrity of the content of the paper

Cite this article: Frithioff A, Frendø M, von Buchwald JH, Trier Mikkelsen P, Sølvsten Sørensen M, Arild Wuyts Andersen S. Automated summative feedback improves performance and retention in simulation training of mastoidectomy: a randomised controlled trial. *J Laryngol Otol* 2022;**136**: 29–36. https://doi.org/10.1017/S0022215121003352

Accepted: 4 February 2021
First published online: 28 October 2021

Key words:
Temporal Bone; Otologic Surgical Procedures

Author for correspondence:
Dr A Frithioff,
Department of Otorhinolaryngology – Head and Neck Surgery, Rigshospitalet, Blegdamsvej 9, Copenhagen Ø DK-2100, Denmark
E-mail: andreasfrit@hotmail.com

Abstract

Objective. This study aimed to investigate the effects of automated metrics-based summative feedback on performance, retention and cognitive load in distributed virtual reality simulation training of mastoidectomy.

Method. Twenty-four medical students were randomised in two groups and performed 15 mastoidectomies on a distributed virtual reality simulator as practice. The intervention group received additional summative metrics-based feedback; the control group followed standard instructions. Two to three months after training, participants performed a retention test without learning supports.

Results. The intervention group had a better final-product score (mean difference = 1.0 points; $p = 0.001$) and metrics-based score (mean difference = 12.7; $p < 0.001$). At retention, the metrics-based score for the intervention group remained superior (mean difference = 6.9 per cent; $p = 0.02$). Also at the retention, cognitive load was higher in the intervention group (mean difference = 10.0 per cent; $p < 0.001$).

Conclusion. Summative metrics-based feedback improved performance and lead to a safer and faster performance compared with standard instructions and seems a valuable educational tool in the early acquisition of temporal bone skills.

Introduction

Most surgical procedures, including temporal bone surgery, require demanding cognitive and psychomotor skills of the surgeon. High-quality training with repeated practice is important to ensure competency, a good surgical outcome and patient safety.[1] Novices have traditionally been introduced to temporal bone surgery through hands-on cadaveric dissection.[2] Nevertheless, because of a decrease in human cadaveric temporal bones available for dissection,[3] interest in alternative training methods such as virtual reality simulation has increased. Even though the evidence for efficacy of virtual reality simulation training is well-established,[4–6] implementation and systematic integration in the curriculum is often limited.[3]

Virtual reality simulation allows the trainee to practice on an unlimited number of cases but also provides the opportunity for directed self-regulated learning.[7] This represents a self-directed learning experience in which the trainees are able to regulate their own learning, scaffolded by instructional design and learning supports provided by the educator and without the presence of a human instructor.[7]

Several benefits of directed self-regulated learning have been reported, such as long-term benefits on performance as well as cost-effectiveness because little or no presence of an instructor is needed.[8,9] Feedback has consistently been identified as a key feature of successful simulation-based surgical training,[10,11] and this can be provided by the simulator itself.[12–15] Altogether, this allows trainees to practice and acquire surgical skills at any time, even at home.[9]

In temporal bone surgical skills training, virtual reality simulation with continuous simulator-integrated tutoring has been found to accelerate the initial learning curves of novices.[16] However, after just a few procedures novices seemingly reach a learning curve plateau because of over-reliance on tutoring.[17] In accordance with the 'guidance hypothesis', this over-reliance on continuous (concurrent) feedback negatively affects performance when the feedback is withdrawn.[18] Feedback also affects the cognitive processes of the learner,[19] and cognitive load theory provides a theoretical framework for understanding learning from a cognitive perspective. The main premise of cognitive load theory is that working memory and information processing capacity is limited, especially for the novice learner.[20] If the sum of cognitive load exceeds the capacity of the learner, this will induce a cognitive overload that negatively affects performance and learning.[21,22]

© The Author(s), 2021. Published by Cambridge University Press

However, some cognitive load (the germane load) is required for the formation of mental schemata (i.e. learning), and continuous feedback can interact with this process.[23]

In contrast to continuous feedback, the use of summative (terminal) feedback appears to result in better learning.[19] In virtual reality temporal bone surgical simulation, such summative feedback has mostly been based on experts' rating performance using structured assessment tools.[24] This is time-consuming and either requires instructor presence during the training situation or later assessment based on recording of the procedure or evaluation of the final product. This makes timely summative feedback nearly impossible. Many simulator-gathered metrics for performance have been suggested,[25] and recent efforts on integrating these into valid assessment enables automated and immediate summative feedback.[14] For other procedural skills such as endoscopy[26] and ultrasound,[27] automatised simulator-based feedback has shown positive effects on novices' performance.

Very little is known about the effects of using summative feedback in virtual reality temporal bone simulation training, but we hypothesise that it will improve end-of-training performance, increase retention of skills and modify cognitive load for the novice. In this study, we therefore want to compare summative feedback based on simulator metrics against standard training without summative feedback in distributed virtual reality simulation training of mastoidectomy.

Materials and methods

Study design, participants and setting

This was a prospective, controlled, randomised trial of an educational intervention. In order to represent true novice trainees, 27 medical students were recruited from the University of Copenhagen, Denmark, and 24 completed the training programme. Figure 1 shows the Consolidated Standards of Reporting Trials flow diagram. Participants were recruited from both clinical and non-clinical semesters, but none had any clinical exposure to temporal bone surgery as this is not part of the pre-graduate curriculum. Prior temporal bone surgical simulation training was the only exclusion criterion. Participants were volunteers and did not receive compensation, and the training was considered an extracurricular activity. The trial took place at the Simulation Centre at Copenhagen Academy of Medical Education and Simulation from October to December 2019 with retention testing in February–March 2020.

Simulation equipment

The virtual reality simulation platform used was an experimental version of the Visible Ear Simulator (version 2.1) that features a range of simulator-integrated metrics for feedback.[14] The Visible Ear Simulator is a high-fidelity virtual reality temporal bone surgical simulator offered as academic freeware online.[28] The simulator uses the Geomagic Touch haptic device (3D Systems, Rockhill, USA) for drilling of a virtual temporal bone with force feedback.

Randomisation

Participants were randomised by the first author (AF) with a 1:1 allocation ratio into two groups using an online random sequence generator before starting the training programme. Upon dropout of one participant, a new participant was recruited and assigned to the same group as the participant who dropped out.

Intervention

Participants in both groups first completed a background questionnaire. Next, participants were introduced to the simulator's navigation and controls by a brief and individual hands-on exercise (5 minutes).

Both training programmes (control and intervention) consisted of five blocks of distributed training: each block was spaced by at least one week and consisted of three identical procedures (complete anatomical mastoidectomy procedures with posterior tympanotomy). As a warm-up, participants were guided by colour-coding (green-lighting) of the bone volume to be drilled in procedure 1 (baseline) but not during any of the following procedures (procedures 2 to 15). Both groups had access to an on-screen, step-by-step dissection guide (standard instructions), which was available at all times during all training procedures. There was no time limit for the procedures.

In contrast to the control group, the intervention group received structured, written summative feedback based on simulator metrics immediately after each procedure.[14] This scoring and feedback sheet (Appendix 1) provides the participant with an overall metrics-based score as well as feedback on choice of drill, bone volume removed, and collisions with important anatomical structures including the dura, facial nerve, chorda tympani, semi-circular canals and the ossicles.

Two months after finishing the initial training, all participants were invited back for retention testing. This consisted of two procedures (procedures 16 and 17) identical to the training procedures but without access to the on-screen instructions and without summative feedback or access to prior scoring sheets.

Outcomes

The primary outcome was manual assessment of the mastoidectomy performance (final-product score). This was done after the trial using a 26-item modified Welling scale for final-product analysis[29] of the end results of the drilling (Figure 2). Two experienced raters (SAWA and MSS), who were blinded to participant, procedure number and group assignment, assessed the performances.

A secondary outcome was the metrics-based score, which is based on five sub-scores combining different metrics and reflecting a correct use of drills, efficiency and goal-directed drilling behaviour. A proficiency level (i.e. pass) for this score has previously been established at a metrics-based score of 83.6 per cent.[14] We further added a collisions score based on the number of collisions with critical structures and also recorded the time used for the procedure.

Cognitive load was another secondary outcome and was measured by secondary-task reaction time, which is an established method for estimating cognitive load.[30] This was done using a reaction timer (American Educational Products, Fort Collins, USA) measuring the time (in 1/100 seconds) it takes to press on a foot switch in response to a beep. Measurements were performed in series of four at baseline (before and after training) and at 5 minutes and 15 minutes during the simulation. Cognitive load was calculated as the mean reaction time during simulation divided by the mean reaction time at baseline (i.e. the relative reaction time).[31]

https://doi.org/10.1017/S0022215121003352 Published online by Cambridge University Press

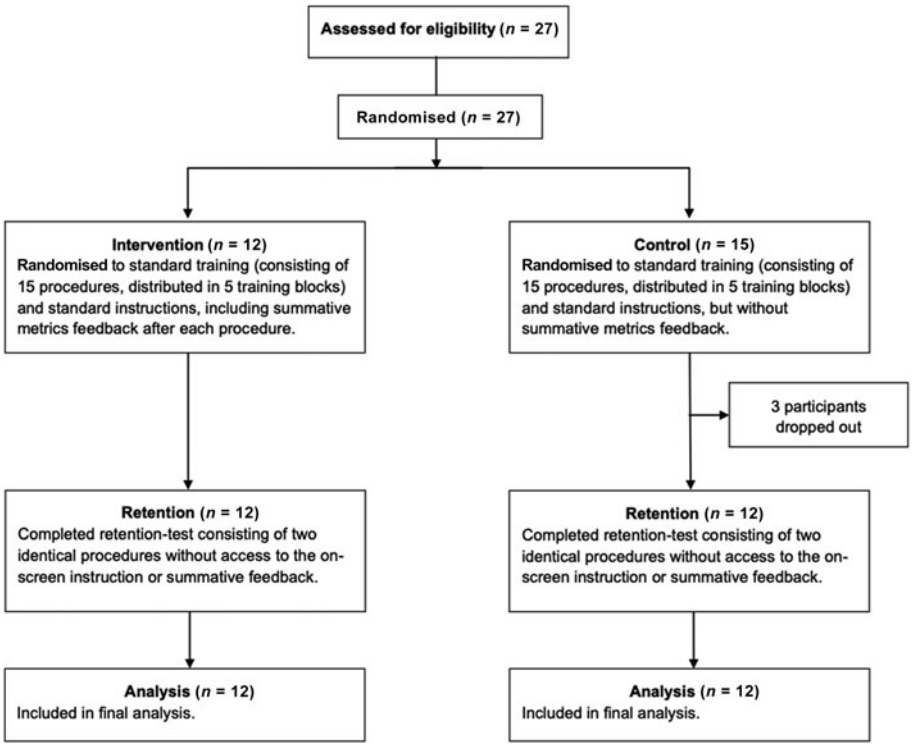


Fig. 1. Consolidated Standards of Reporting Trials flow diagram.

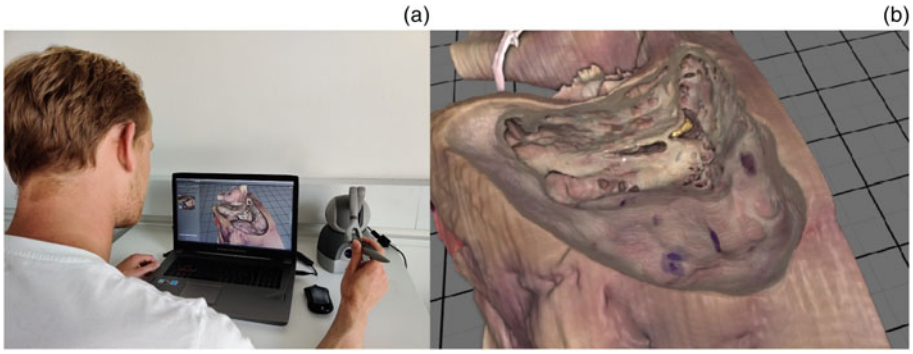


Fig. 2. Simulation set-up (a), and an example of a mastoidectomy final product after a training procedure (b).

Sample size

Sample size calculations were based on experience from similar studies because sample size calculations for repeated measurement designs are not well-defined. Therefore, we chose 12 participants in each arm which, based on previous studies, should be able to detect a 10 per cent difference in the final-product outcome.

Statistical methods

Data were analysed using SPSS® (version 25) for Mac® OSX. Because of repeated measurements, linear mixed models, using the principles outlined by Leppink,[32] were used in the analyses. Models were iteratively built to investigate the different factors and their interactions as fixed effects: for the final-product score, the final model included group, procedure number and rater; for the metrics-based score outcomes, the final model included group and the procedure number; for the cognitive load outcome, the final model included only group as timing of reaction because time measurement during the procedure (at 5 minutes and 15 minutes) and procedure number was not found to influence cognitive load; for the retention procedures, the corresponding models included group and rater (final-product score) or group only (metrics-based score and cognitive load). Estimated marginal means and *p*-values of the linear mixed models were reported. *P*-values less than 0.05 were considered statistically significant.

Ethics

The regional ethical committee of the Capital Region of Denmark found this educational trial exempt (H-19069755). Written consent was obtained from participants.

Results

Participants in the control and intervention groups had similar baseline characteristics including self-reported computer skills and gaming frequency (Table 1).

Effects on final-product score

For the expert assessment of the final-product score performance, the two groups had similar performance at baseline (i.e. the warm-up procedure) (mean difference = 0.7 points; $p = 0.45$). During the trial, final-product score increased with repeated practice (0.08 points per procedure; $p = 0.045$) in

https://doi.org/10.1017/S0022215121003352 Published online by Cambridge University Press

Table 1. Participant characteristics

Parameter	Intervention: structured summative feedback	Control: no feedback
Participants (*n*)	12	12
Age (mean (SD; years)	23 (1.4)	26 (9.9)
Sex (*n*)		
– Female	8	6
– Male	4	6
Weekly computer usage excluding work hours (mean (SD); hours)	8.1 (6.8)	9.8 (5.3)
Self-reported computer skills (mean (SD); Likert scale 1–7)	5.1 (0.7)	4.3 (1.4)
Gaming frequency (mean (SD); Likert scale 1–5)	3.9 (1.1)	3.6 (1.4)

SD = standard deviation

Table 2. Performance in retention testing

Group	Mean final-product score (points (95% CI))	Mean metrics-based score (% (95% CI))	Relative increase in cognitive load (% (95% CI))
Intervention	15.3 (14.2 to 16.4)	81.5 (77.5 to 85.4)	30.0 (26.4 to 33.5)
Control	14.4 (13.3 to 15.4)	74.6 (70.6 to 78.6)	20.0 (16.5 to 23.6)
P-value	0.23	0.02	<0.001

CI = confidence interval

both groups as expected (Figure 3). Importantly, we found that the intervention group significantly outperformed the control group (mean difference = 1.0 point; $p = 0.001$). At retention testing, the intervention group performed slightly better than the control group, but this was not statistically significant (Table 2).

Effects on metrics-based score, collisions and time

For performance assessment using the automated metrics-based score, we found similar results. Participants scored similarly at baseline (mean difference = 1.9; $p = 0.60$) and repeated practice increased the metrics-based score (1.6 per cent per procedure; $p < 0.001$). During training, the intervention group performed far superiorly to the control group (mean difference = 12.7 per cent; $p < 0.001$; Figure 4). This also resulted in the intervention group having more total performances that passed the pre-defined proficiency level compared with the control group (41.6 per cent *vs* 8.8 per cent; $p < 0.001$). Finally, at retention testing, the intervention group continued to have a higher metrics-based score compared with the control-group (mean difference = 6.9 per cent; $p = 0.02$) (Table 2). We found a poor correlation between the metrics-based score and final-product score ($r^2 = -0.04$).

For collisions and time, the intervention group made significantly fewer total collisions (mean, 43.4 *vs* 54.1; $p < 0.001$) and also completed the procedure using less time compared with the control group (mean difference = 4.6 minutes; $p < 0.001$). At retention testing, we found no statistically significant difference in the number of collisions (mean difference = 6.3; $p = 0.31$) or time (mean difference = 2.4 minutes; $p = 0.35$).

Effects on cognitive load

There was no difference in cognitive load between the intervention and control group at baseline (mean difference = 6.2 per cent; $p = 0.33$) or during training (mean difference =1 per cent; $p = 0.20$), and cognitive load did not decrease with repeated practice. In contrast, the intervention group was found to have a higher cognitive load compared with the control group during retention testing (mean difference = 10 per cent; $p < 0.001$) (Table 2). When comparing cognitive load at the end of training (procedures 13–15) with the retention test (procedures 16–17), cognitive load was 7.1 per cent higher for the intervention group ($p = 0.005$) whereas the control group experienced a 1.8 per cent decrease in cognitive load ($p = 0.005$).

Discussion

Overall, we found that the summative feedback intervention improved novices' performances during virtual reality simulation training considerably and accelerated the initial learning curve using both manual assessment and automated scoring based on simulator-metrics as the outcome. Further, the intervention resulted in fewer collisions with key structures (e.g. the facial nerve) and also decreased time to complete the procedure. At the retention test, metrics-based score remained higher for the intervention group; however, there was no significant difference in performance for the final-product score. The intervention did not affect cognitive load during training; however, during the retention testing, the cognitive load induced in the intervention group was significantly increased.

It is not surprising that the intervention group had a higher metrics-based score compared with the control group during the training because the intervention group received this score along with feedback based on the same metrics after each completed procedure. However, the control group did not receive any summative feedback. The learning curves of both groups (Figure 3 and 4) follow a classic pattern with initial fast acceleration of performance followed by gradual plateauing after just a few procedures (i.e. negatively accelerated learning curves).[16] The difference in metrics-based score between groups at procedure two reflects the feedback given to the intervention group received after their warm-up procedure (procedure one).

The metrics-based score mainly reflects process and efficiency, such as choosing the appropriate burr size and type, time aspects, and goal-directed behaviour. In line with previous studies,[14] we found the metrics-based score to correlate poorly with the manual final-product score, which considers only the end result and emphasises safety-related parts of the procedure, such as avoiding drilling holes and damaging key structures.[14,33] Nevertheless, providing the participants with the summative metrics-based score and collision information had a positive impact on their final-product performance (final-product score). Consequently, the automated summative feedback appears to be a strong educational tool for directed, self-regulated learning. Ultimately, this allows learners to develop basic surgical skills in mastoidectomy, reducing the need for human instructors[7] who can be saved for more advanced training, such as on cadavers.

https://doi.org/10.1017/S0022215121003352 Published online by Cambridge University Press

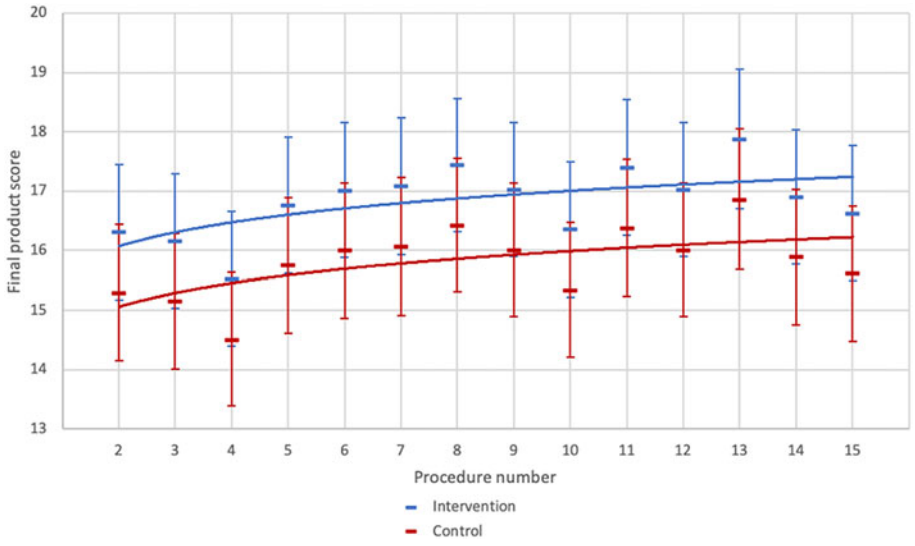


Fig. 3. Means plot (estimated marginal means) of final-product score learning curves of training sessions (procedures 2–15). Bars mark 95 per cent confidence interval. The first procedure (i.e. warm-up) was not included in the figure as participants were guided by colour-coding (green-lighting) of the bone volume to be drilled in the procedure.

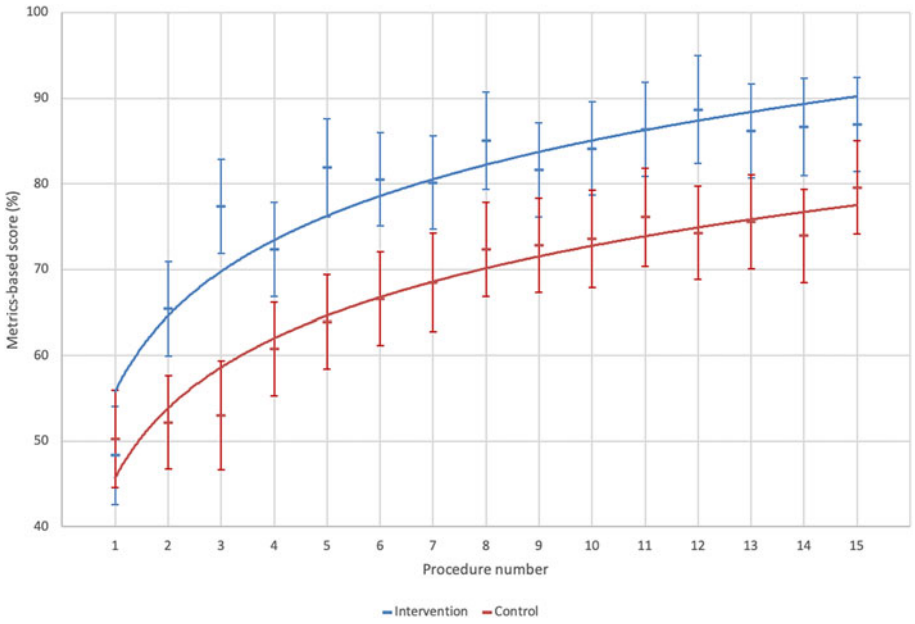


Fig. 4. Means plot (estimated marginal means) of metrics-based score learning curves of training sessions (procedures 1–15). Bars mark 95 per cent confidence interval. The intervention group received summative metrics-based feedback after the first procedure (i.e. warm-up); this results in an immediate difference in metrics-based score between the two groups in procedure two.

Our study adds new knowledge for several reasons: first, it is the first study to investigate automated summative feedback in temporal bone training because all previous studies have used continuous feedback (real-time feedback), through green-lighting for example.[12,34,35] Next, we have studied the effect in a prolonged, distributed training programme, which is closer to real-life training conditions. Also, we included retention testing after two to three months to study the effect on longer term performance. Finally, we did not only measure the performance as simulator-gathered score (metrics-based score), but also as assessed by experts using an established mastoidectomy assessment tool (final-product score).

This study on summative feedback was motivated by previous findings, which demonstrated that real-time feedback may have negative effects when it is withdrawn.[16] This is likely explained by tutoring over-reliance, which easily occurs in early stages of learning. In contrast, we now report how summative feedback does not have the same negative impact on acquisition of skills or retention, which is consistent with 'the guidance hypothesis'.[19] We cannot, however, conclude an ideal number of procedures with summative feedback, where performance remains stable after withdrawal of the feedback. A future step would be to further investigate the effects of summative feedback on transfer of simulation skills to performance in cadaveric dissection.

- Simulation-based training can be used for self-directed acquisition of temporal bone surgical skills
- Automated feedback is key for effective directed, self-regulated learning, but the best way to provide such feedback is unknown
- Metrics-based summative feedback leads to a more efficient and safer drilling behaviour in virtual reality mastoidectomy training
- Metrics-based summative feedback is a strong educational tool for novices in the early acquisition of temporal bone surgical skills
- Metrics-based summative feedback can be integrated as an automated learning support in simulation-based temporal bone training

We found cognitive load to be similar and stable for the two groups during training. Surprisingly, during the retention testing, a higher cognitive load was induced in the intervention group. Other studies within virtual reality simulation-based training of mastoidectomy have found that other learning supports affect cognitive load[36,37]: for example, continuous

https://doi.org/10.1017/S0022215121003352 Published online by Cambridge University Press

feedback through automated tutoring reduces cognitive load during training but at the cost of inducing a very high cognitive load when tutoring is withdrawn. According to cognitive load theory, a low cognitive load during training of complex skills is not unconditionally beneficial for actual learning because some cognitive resources need to be allocated for the learning process itself.[38,39] The sub-components of cognitive load are difficult to measure separately and because relative reaction time estimates the total cognitive load, we are not able to determine if there are differences in the distribution between sub-components in our two groups.

A limitation of our study is that we used medical students as participants. In contrast to even first-year residents, medical students are true novices in relation to the procedure, and their learning objectives and motivation might therefore be very different. Consequently, we cannot directly extrapolate our results to more experienced learners, and future studies should elucidate whether our findings also apply to experienced learners or other specialties. Furthermore, we did not investigate a transfer outcome such as performance in cadaver dissection or in the operating theatre. As the virtual reality environment differs from the operating theatre in several ways (e.g. no bleeding or need for suctioning), a complete transfer of skills cannot be expected.[5,40,41] A strength of our study is that our training programme was distributed (i.e. comprised multiple sessions separated by several days), which is not only an important part of directed self-regulated learning[40] but also results in better acquisition of skills in temporal bone surgery compared with massed practice.[16,41] Validity evidence for the metrics-based score that we used for summative feedback has been established.[14] However, metrics are simulator-specific and vary between simulators,[25] and consequently, integration of metrics-based score for summative feedback in other simulators requires context-specific validity evidence to be collected.

Our study has several implications for virtual reality simulation-based training in temporal bone surgery. Automated, summative metrics-based feedback leads to an improved training and retention performance, supporting directed, self-regulated learning where the trainee can practice without the presence of human instructors. Furthermore, learning curves were accelerated, and even though the performance-gap between the control and intervention group in this study might diminish over time, summative metrics-based feedback can help reduce training time to reach a certain level of competence. The metrics-based feedback also resulted in a more efficient and safer drilling behaviour, which hopefully could translate into a safe clinical behaviour as well. Finally, virtual reality simulation training should be considered a first step before using other training modalities, saving cadaver and instructional resources, for example, until the trainee has demonstrated adequate skills in simulation. A comprehensive surgical training curriculum should integrate different training modalities and implement mastery learning where feedback, score-tracking and testing constitute crucial elements.[42]

Conclusion

Summative metrics-based feedback has several positive effects on novices' performance in virtual reality simulation-based training of temporal bone surgery. This includes increasing performance during training, reducing the number of collisions with key structures and reducing time for each simulated procedure. These positive effects seemed to be retained to some degree after two to three months. For these reasons, summative feedback can potentially lead to a safer, better and more efficient performance. The intervention did not seem to affect the total cognitive load during training, most likely because cognitive resources were allocated towards germane load (i.e. formation of mental schemata). Altogether, automated metrics-based summative feedback is a valuable educational tool in novices' initial mastoidectomy skills acquisition and can be integrated as a support for directed, self-regulated learning in the basic temporal bone skills training curriculum.

Acknowledgements. Steven Andersen has received research funding for his postdoctoral study from the Independent Research Fund Denmark (8026-00003B). The remaining authors have no other sources of funding or support to declare.

Competing interests. None declared

References

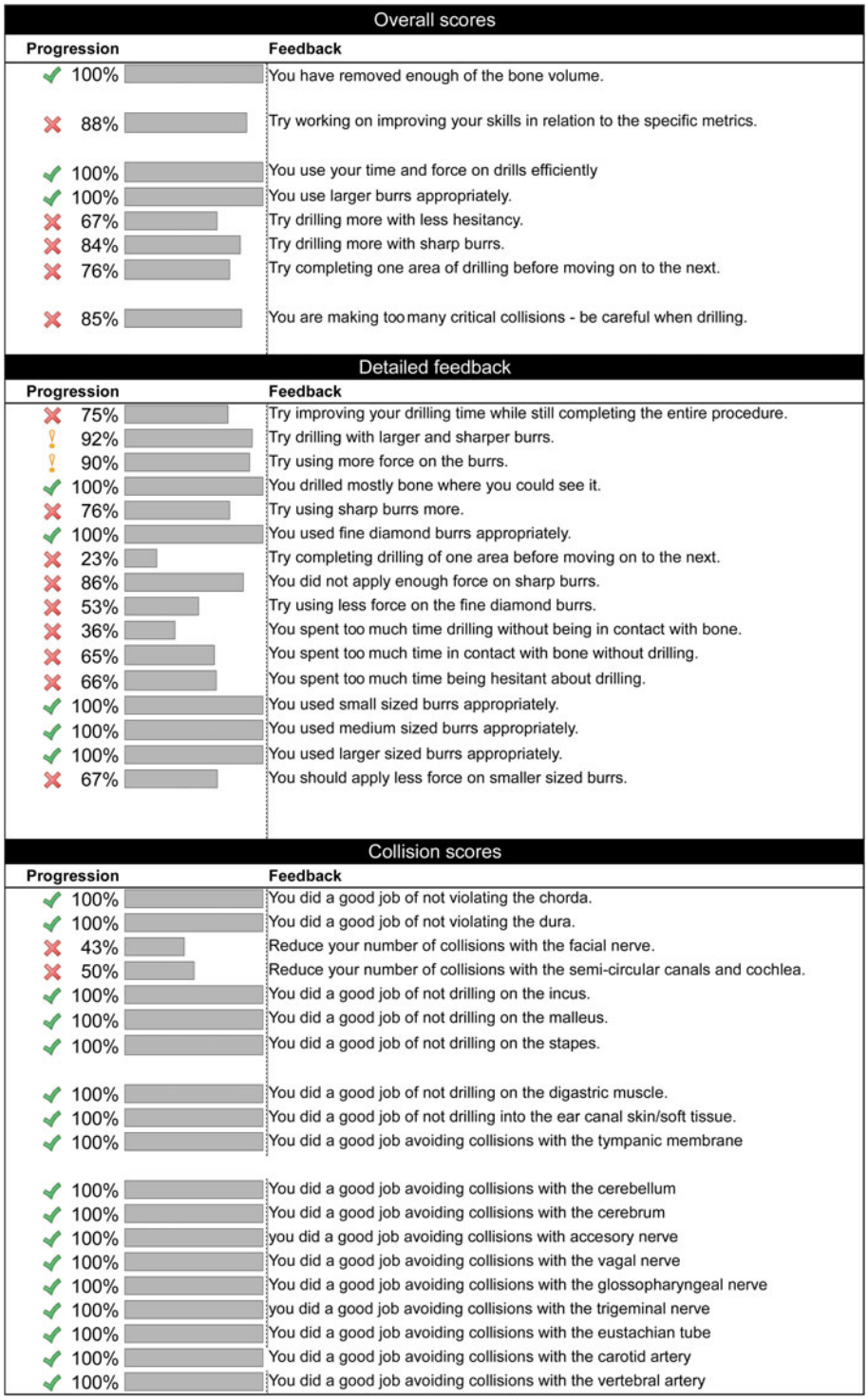
1 Reznick RK. Teaching and testing technical skills. *Am J Surg* 1993;**165**:358–61
2 George AP, De R. Review of temporal bone dissection teaching: how it was, is and will be. *J Laryngol Otol* 2010;**124**:119–25
3 Frithioff A, Sørensen MS, Andersen SAW. European status on temporal bone training: a questionnaire study. *Eur Arch Otorhinolaryngol* 2018;**275**:357–63
4 Zhao YC, Kennedy G, Yukawa K, Pyman B, O'Leary S. Improving temporal bone dissection using self-directed virtual reality simulation: results of a randomized blinded control trial. *Otolaryngol Head Neck Surg* 2011;**144**:357–64
5 Andersen SAW, Foghsgaard S, Konge L, Cayé-Thomasen P, Sørensen MS. The effect of self-directed virtual reality simulation on dissection training performance in mastoidectomy. *Laryngoscope* 2016;**126**:1883–8
6 Javia L, Deutsch ES. A systematic review of simulators in otolaryngology. *Otolaryngol Head Neck Surg* 2012;**147**:999–1011
7 Brydges R, Dubrowski A, Regehr G. A new concept of unsupervised learning: directed self-guided learning in the health professions. *Acad Med* 2010;**85**:S49–55
8 Brydges R, Nair P, Ma I, Shanks D, Hatala R. Directed self-regulated learning versus instructor-regulated learning in simulation training. *Med Educ* 2012;**46**:648–56
9 Frendø M, Thingaard E, Konge L, Sølvsten M, Steven S. Decentralized virtual reality mastoidectomy simulation training: a prospective, mixed-methods study. *Eur Arch Otorhinolaryngol* 2019;**276**:2783–9
10 Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence. *Acad Med* 2013;**88**:872–83
11 Issenberg SB, Mcgaghie WC, Petrusa ER, Gordon DL, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach* 2005;**27**:10–28
12 Wijewickrema S, Zhou Y, Ioannou I, Copson B, Piromchai P *et al.* Presentation of automated procedural guidance in surgical simulation: results of two randomised controlled trials. *J Laryngol Otol* 2018;**132**:257–63
13 Kerwin T, Stredney D, Wiet GJ, Shen HW. Virtual Mastoidectomy performance evaluation through multi-volume analysis. *Computer Assist Radiol Surg* 2013;**8**:51–61
14 Andersen SAW, Mikkelsen PT, Sørensen MS. Expert sampling of VR simulator metrics for automated assessment of mastoidectomy performance. *Laryngoscope* 2019;**129**:2170–7
15 Zirkle M, Roberson DW, Leuwer R, Dubrowski A. Using a virtual reality temporal bone simulator to assess otolaryngology trainees. *Laryngoscope* 2007;**117**:258–63
16 Andersen SAW, Konge L, Cayé-Thomasen P, Sørensen MS. Learning curves of virtual mastoidectomy in distributed and massed practice. *JAMA Otolaryngol Head Neck Surg* 2015;**141**:913–18
17 Andersen SAW, Konge L, Mikkelsen PT. Mapping the plateau of novices in virtual reality simulation training of mastoidectomy. *Laryngoscope* 2017;**127**:907–14
18 Park J, Shea CH, Wright DL, Shea CH, Reduced-frequency DLW, Park J *et al.* Reduced-frequency concurrent and terminal feedback: a test of the guidance hypothesis. *J Mot Behav* 2000;**32**:287–96
19 Hatala, R, Cook, D, Zendajas, B, Hamstra, S, Brydges R. Feedback for simulation-based procedural skills training: a meta-analysis and critical narrative synthesis. *Adv Heal Sci Educ Theory Pract* 2014;**19**:251–72

https://doi.org/10.1017/S0022215121003352 Published online by Cambridge University Press

20 Sweller J. Cognitive load during problem solving: effects on learning. *Cognitive Science* 1988;**12**:257–85
21 Haji FA, Cheung JJH, Woods N, Regehr G, de Ribaupierre S, Dubrowski A. Thrive or overload? The effect of task complexity on novices' simulation-based learning. *Med Educ* 2016;**50**:955–68
22 Frithioff A, Frendø M, Mikkelsen PT, Sørensen MS, Andersen SAW. Ultra-high-fidelity virtual reality mastoidectomy simulation training: a randomized, controlled trial. *Eur Arch Otorhinolaryngol* 2020;**277**:1335–41
23 Van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: design principles and strategies. *Med Educ* 2010;**44**:85–93
24 Sethia R, Kerwin TF, Wiet GJ. Performance assessment for mastoidectomy. *Otolaryngol Head Neck* Surg. 2017;**156**:61–69
25 Al-Shahrestani F, Sørensen MS, Andersen SAW. Performance metrics in mastoidectomy training: a systematic review. *Eur Arch Otorhinolaryngol* 2019;**276**:657–64
26 Vilmann AS, Norsk D, Bo M, Svendsen S. Computerized feedback during colonoscopy training leads to improved performance: a randomized trial. *Gastrointest Endosc* 2018;**88**:869–76
27 Ahmed OMA, Niessen T, Gallagher AG, Breslin DS, Dunngalvin A, Shorten GD. The effect of metrics-based feedback on acquisition of sonographic skills relevant to performance of ultrasound-guided axillary brachial plexus block. *Anaesthesia* 2017;**9**:1117–24
28 Sørensen MS, Mosegaard J, Trier P. The visible ear simulator: a public PC application for GPU-accelerated haptic 3D simulation of ear surgery based on the visible ear data. *Otol Neurotol* 2009;**30**:484–7
29 Andersen SAW, Cayé-Thomasen P, Sørensen MS. Mastoidectomy performance assessment of virtual simulation training using final-product analysis. *Laryngoscope* 2015;**125**:431–5
30 Naismith LM, Cavalcanti RB. Validity of cognitive load measures in simulation-based training: a systematic review. *Acad Med* 2015;**90**:24–35
31 Andersen SAW, Mikkelsen PT, Konge L, Cayé-Thomasen P, Sørensen MS. Cognitive load in mastoidectomy skills training: virtual reality simulation and traditional dissection compared. *J Surg Educ* 2016;**73**:45–50
32 Leppink J. Data analysis in medical education research: a multilevel perspective. *Perspect Med Educ* 2015;**4**:14–24
33 Zirkle M, Taplin MA, Anthony R, Dubrowski A. Objective assessment of temporal bone drilling skills. *Ann Otol Rhinol Laryngol* 2007;**116**:793–8
34 Davaris M, Wijewickrema S, Zhou Y, Piromchai P, Bailey J, Kennedy G *et al*. The importance of automated real-time performance feedback in virtual reality temporal bone surgery training. In: Isotani S, Millán E, Ogan A, Hastings P, McLaren B, Luckin R, eds. *Artificial Intelligence in Education*. Cham: Springer International Publishing, 2019;96–109
35 Wijewickrema S, Piromchai P, Zhou Y, Ioannou I, Bailey J, Kennedy G *et al*. Developing effective automated feedback in temporal bone surgery simulation. *Otolaryngol Head Neck Surg* 2015;**152**:1082–8
36 Andersen SAW, Mikkelsen PT, Sørensen MS. The Effect of simulator-integrated tutoring for guidance in virtual reality simulation training. *Simul Healthc* 2020;**15**:147–53
37 Andersen SAW, Frendø M, Guldager M, Sørensen MS. Understanding the effects of structured self-assessment in directed, self-regulated simulation-based training of mastoidectomy: a mixed methods study. *J Otol* 2020;**15**:117–23
38 Andersen SAW, Frendø M, Sørensen MS. Effects on cognitive load of tutoring in virtual reality simulation training. *MedEdPublish* 2020;**9**:1–6
39 Van Merriënboer JJG, Kester L, Pass F. Teaching complex rather than simple tasks: balancing intrinsic and germane load to enhance transfer of learning. *Appl Cogn Psychol* 2006;**20**:343–52
40 Gawecki W, Wegrzyniak M, Mickiewicz P, Talar M, Wierzbicka M, Gawłowska MB. The impact of virtual reality training on the quality of real antromastoidectomy performance. *J Clin Med*. 2020;**9**:3197
41 Andersen SAW, Foghsgaard S, Cayé-Thomasen P, Sørensen MS. The effect of a distributed virtual reality simulation training program on dissection mastoidectomy performance. *Otol Neurotol* 2018;**39**:1277–84
42 Smith S, Lonie J. Mastery learning: how is it helpful? An analytical review. *Adv Med Educ Pract* 2017;**8**:269–75

https://doi.org/10.1017/S0022215121003352 Published online by Cambridge University Press

Appendix 1. Metrics-based scoring sheet provided after each procedure



Overall scores

Progression	Feedback
✓ 100%	You have removed enough of the bone volume.
✗ 88%	Try working on improving your skills in relation to the specific metrics.
✓ 100%	You use your time and force on drills efficiently
✓ 100%	You use larger burrs appropriately.
✗ 67%	Try drilling more with less hesitancy.
✗ 84%	Try drilling more with sharp burrs.
✗ 76%	Try completing one area of drilling before moving on to the next.
✗ 85%	You are making too many critical collisions - be careful when drilling.

Detailed feedback

Progression	Feedback
✗ 75%	Try improving your drilling time while still completing the entire procedure.
! 92%	Try drilling with larger and sharper burrs.
! 90%	Try using more force on the burrs.
✓ 100%	You drilled mostly bone where you could see it.
✗ 76%	Try using sharp burrs more.
✓ 100%	You used fine diamond burrs appropriately.
✗ 23%	Try completing drilling of one area before moving on to the next.
✗ 86%	You did not apply enough force on sharp burrs.
✗ 53%	Try using less force on the fine diamond burrs.
✗ 36%	You spent too much time drilling without being in contact with bone.
✗ 65%	You spent too much time in contact with bone without drilling.
✗ 66%	You spent too much time being hesitant about drilling.
✓ 100%	You used small sized burrs appropriately.
✓ 100%	You used medium sized burrs appropriately.
✓ 100%	You used larger sized burrs appropriately.
✗ 67%	You should apply less force on smaller sized burrs.

Collision scores

Progression	Feedback
✓ 100%	You did a good job of not violating the chorda.
✓ 100%	You did a good job of not violating the dura.
✗ 43%	Reduce your number of collisions with the facial nerve.
✗ 50%	Reduce your number of collisions with the semi-circular canals and cochlea.
✓ 100%	You did a good job of not drilling on the incus.
✓ 100%	You did a good job of not drilling on the malleus.
✓ 100%	You did a good job of not drilling on the stapes.
✓ 100%	You did a good job of not drilling on the digastric muscle.
✓ 100%	You did a good job of not drilling into the ear canal skin/soft tissue.
✓ 100%	You did a good job avoiding collisions with the tympanic membrane
✓ 100%	You did a good job avoiding collisions with the cerebellum
✓ 100%	You did a good job avoiding collisions with the cerebrum
✓ 100%	you did a good job avoiding collisions with accesory nerve
✓ 100%	You did a good job avoiding collisions with the vagal nerve
✓ 100%	You did a good job avoiding collisions with the glossopharyngeal nerve
✓ 100%	you did a good job avoiding collisions with the trigeminal nerve
✓ 100%	You did a good job avoiding collisions with the eustachian tube
✓ 100%	You did a good job avoiding collisions with the carotid artery
✓ 100%	You did a good job avoiding collisions with the vertebral artery

https://doi.org/10.1017/S0022215121003352 Published online by Cambridge University Press