Full citation: 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 Jan;136(1):29-36.
DOI: 10.1017/S0022215121003352

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

Authors: Andreas Frithioff, MD[1,2], Martin Frendø, MD[1,2], Josefine Hastrup Buchwald, BSc[1,2], Peter Trier Mikkelsen, MSc[3], Mads Sølvsten Sørensen, MD, DMSc[1], Steven Arild Wuyts Andersen, MD, PhD[1,2]

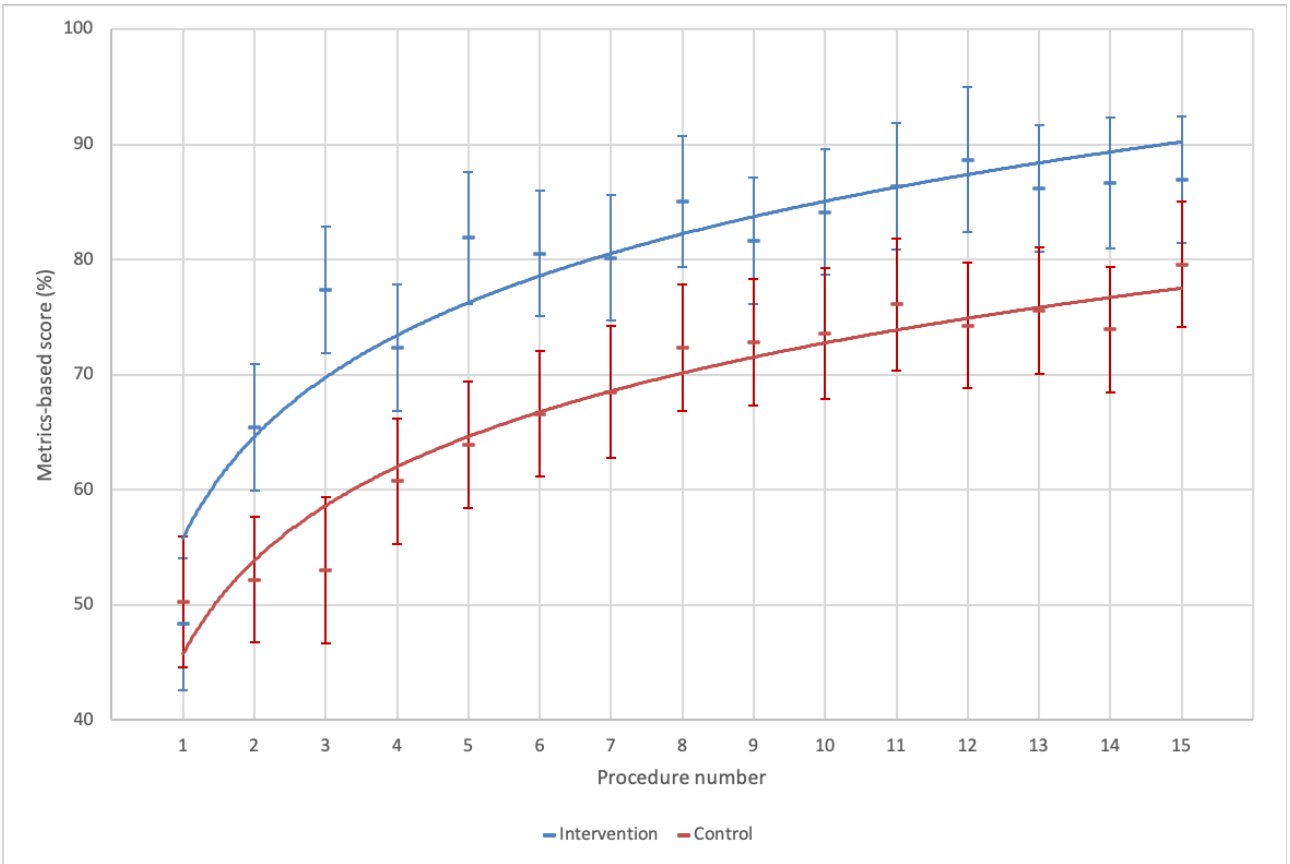
Affiliations:

1. Dept. of Otorhinolaryngology—Head & Neck Surgery, Rigshospitalet, Copenhagen, Denmark.
2. Copenhagen Academy for Medical Education and Simulation (CAMES), The Capital Region of Denmark, Copenhagen, Denmark.
3. The Alexandra Institute, Aarhus, Denmark.

Correspondence: Andreas Frithioff, MD. Department of Otorhinolaryngology—Head & Neck Surgery, Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen Ø, Denmark. Phone: 0045 35452071. E-mail: Andreasfrit@hotmail.com.

Funding: 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.



Abstract

Objective: To investigate the effects of automated metrics-based summative feedback on performance, retention and cognitive load in distributed virtual reality (VR) simulation training of

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.

Keywords: Temporal bone surgery; mastoidectomy; surgical education; simulation-based training; summative feedback; directed self-regulated learning

Introduction

Most surgical procedures—including in 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, due to a decrease of human cadaveric temporal bones available for dissection (3), interest in alternative training methods such as virtual reality (VR) simulation has increased. Even though evidence for efficacy of VR 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 (DSRL) (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 DSRL have been reported for example 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, VR simulation with continuous simulator-integrated tutoring has been found to accelerate the initial learning curves of novices (16). However, after just few procedures novices reach a seeming plateau of learning curve due to tutoring over-reliance (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). 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 VR 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). In other procedural skills such as endoscopy (26) and ultrasound (27), automatized simulator-based feedback has shown positive effects on novices' performance.

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

Material and methods

Study design, participants and setting

This was a prospective, controlled, randomized trial of an educational intervention.

To represent true novice trainees, twenty-seven medical students were recruited from the University of Copenhagen, Denmark and twenty-four completed the training program. Figure 1 shows the CONSORT 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 (CAMES) from October–December 2019 with retention testing in February–March 2020.

Simulation equipment

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

Randomization

Participants were randomized by the first author (A.F.) with a 1:1 allocation ratio into two groups using an online random sequence generator before starting the training program. Upon dropout, a new participant was recruited and assigned the same group as the dropped-out participant.

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 programs (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 mastoidectomies with posterior tympanotomy). As a warm-up, participants were guided by color-coding (green-lighting) of the bone volume to be drilled in procedure 1 (baseline) but not during any of the following procedures (procedures 2–15). 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 A, Supplemental Digital Content) provides the participant an overall metrics-based score as well as feedback on choice of drill, bone volume removed, 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 (procedure 16–17) identical to the training procedures, however, 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, FPS). 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 (S.A. and M.S.), who were blinded to participant, procedure number, and group assignment, assessed the performances.

A secondary outcome was the metrics-based score (MBS), which is based on five sub-scores combining different metrics and reflecting a correct use of drills, efficiency, and goal-directed drilling behavior. A proficiency level (i.e. pass) for this score has previously been established at a MBS of 83.6% (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 (CL) was another secondary outcome and was measured by secondary-task reaction time, which is an established method for estimating CL (30). This was done using reaction timer (American educational products, LLC, USA) measuring the time (in 1/100 s) it takes to press on a foot switch in response to at beep. Measurements were performed in series of four at baseline (before and after training) and at t=5 min and t=15 min 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).

Sample size

Sample size calculations were based on experience from similar studies because sample size calculations for repeated measurements 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% difference in the final-product outcome.

Statistical methods

Data were analyzed using SPSS version 25 (IBM, Armonk, NY, USA) for Mac OSX. Due to repeated measurements, linear mixed models (LMM) 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 FPS, the final model included group, procedure number, and rater; for the MBS outcomes, the final model included group and the procedure number; for the CL outcome, the final model included only group as timing of reaction because time measurement during the procedure (t=5 min and t=15 min) and procedure number was not found to influence CL; for the retention procedures, the corresponding models included group and rater (FPS) or group only (MBS and CL). Estimated marginal means and *P* values of the LMM are reported. *P* values <0.05 are 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, participant demographics).

Effects on final-product score (FPS)

For the expert assessment of the FPS 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, FPS increased with repeated practice (0.08 points per procedure; p=0.045) in both groups as expected (Figure 3). Importantly, we found that the intervention group significantly outperformed the control group (mean diff.=1.0 points, 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 (MBS), collisions and time

For performance assessment using the automated MBS, we found similar results. Participants scored similarly at baseline (mean diff.=1.9; p=0.60) and repeated practice increased the MBS (1.6% per procedure; p<0.001). During training, the intervention group performed far superiorly to the control group (mean difference 12.7%; 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% vs. 8.8%; p<0.001). Finally, at retention testing, the intervention group continued to have a higher MBS compared with the control-group (mean diff.=6.9%; p=0.02) (Table 2). We found a poor correlation between the MBS and FPS (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 diff.=4.6 min; p<0.001). At retention testing, we found no statistically significant difference in the number of collisions (mean diff.=6.3; p=0.31) or time (mean diff.=2.4 min; p=0.35).

Effects on cognitive load

There was no difference in CL between the intervention and control group at baseline (mean diff.=6.2%; p=0.33) or during training (mean diff.=1%; p=0.20) and CL did not decrease with

repeated practice. In contrast, the intervention group was found to have a higher CL compared with the control group during retention testing (mean diff.=10%; $p<0.001$) (Table 2). When comparing CL at the end-of-training (procedures 13–15) with the retention test (procedures 16–17), CL was 7.1% higher for the intervention group ($p=0.005$) whereas the control group experienced a 1.8% decrease in CL ($p=0.005$)

Discussion

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

It is not surprising that the intervention group had a higher MBS compared with the control group during the training since the intervention group received this score along with feedback based on the same metrics after each completed procedure. The control group however, did not receive any summative feedback. The learning curves of the two groups (Figure 3 & 4) follows a classic pattern with fast acceleration of performance initially and then gradually plateauing after just a few performance (i.e. negatively accelerated learning curve) (16). The difference in MBS between the two groups observed already at procedure number two reflects the feedback the intervention group received after completing the warm-up procedure (procedure one). The MBS mainly reflects process and efficiency for example choosing the appropriate burr size and type, time aspects, and

goal-directed behavior. In line with previously[14], we found the MBS to correlate poorly with the manually FPS, which considers only the end result and emphasizes safety-related parts of the procedure such as avoiding drilling holes and damaging key structures (14,33). Nevertheless, providing the participants with the summative MBS and collision information had a positive impact on their final-product performance (FPS). 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 for example on cadavers.

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

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 consistence with “the guidance hypothesis”(19). A future step would be to investigate further the effects of summative feedback on transfer of simulation skills to performance in cadaveric dissection.

We found cognitive load to be similar and stable for the two groups during training. Surprisingly, during the retention testing, a higher CL was induced in the intervention group. Other studies within VR simulation-based training of mastoidectomy have found that other learning supports affect CL (36,37): For example, continuous feedback through automated tutoring reduces CL during training but at the cost of inducing a very high CL when tutoring is withdrawn. According to cognitive load theory, a low CL during training of complex skills is not unconditionally beneficial for actual learning as indeed some cognitive resources need to be allocated for the learning process itself (38). The sub-components of CL are difficult to measure separately and because relative reaction time estimates the total CL, 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 investigate if the findings also apply to for an example ORL or neurosurgery residents. Further, we did not investigate a transfer outcome such as performance in cadaver dissection or in the OR. As the VR environment differs from the OR in several ways (e.g. no bleeding or need for handling suction exists), a complete transfer of skills cannot be expected. (5,40,41) A strength of our study is that our training program was distributed (i.e. comprised multiple sessions separated by several days), which not only is an important part of DSRL (40) but also results in better acquisition of skills in temporal bone surgery compared with massed practice (16,41). Validity evidence for the MBS we used for summative feedback has been established (14). However, metrics are simulator-specific and vary between simulators (25) and consequently, integration of MBS for summative feedback in other simulators requires context-specific validity evidence to be collected.

Our study has several implications for VR 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. Further, 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 a given level of competence. The metrics-based feedback also resulted in a more efficient and safer drilling behavior, which hopefully could translate into a safe clinical behavior as well. Finally, VR simulation training should be considered a first step before using other training modalities, saving for example cadaver and instructional resources 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 VR 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-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.

References

1. Reznick RK. Teaching and Testing Technical Skills. Surg Educ. 1993;165(March).
2. George AP, De R. Review of temporal bone dissection teaching: How it was, is and will be. J Laryngol Otol. 2010;124(2):119–25.
3. Frithioff A, Sørensen MS, Andersen SAW. European status on temporal bone training: a questionnaire study. Eur Arch Oto-Rhino-Laryngology [Internet]. 2018;275(2):357–63. Available from: http://dx.doi.org/10.1007/s00405-017-4824-0
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(3):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(6):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(10):49–55.
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 Oto-

Rhino-Laryngology [Internet]. 2019; Available from: https://doi.org/10.1007/s00405-019-05572-9

10. Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-Enhanced Simulation to Assess Health Professionals : A Systematic Review of Validity Evidence , Research Methods , and Reporting Quality. 2013;88(6).
11. Issenberg SB, Mcgaghie WC, Petrusa ER, Gordon DL, Scalese RJ, Issenberg SB, et al. Features and uses of high-fidelity medical simulations that lead to effective learning : a BEME systematic review Features and uses of high-fidelity medical simulations that lead to effective learning : a BEME systematic review *. 2009;142–59.
12. Wijewickrema S, Zhou Y, Ioannou I, Copson B, Piromchai P, Yu C. 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 H-W. Virtual Mastoidectomy Performance Evaluation through Multi-Volume Analysis. 2014;8(1):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(10):913–8.
17. Andersen SAW, Konge L, Mikkelsen PT. Mapping the Plateau of Novices in Virtual Reality Simulation Training of Mastoidectomy.
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.

2010;2895.

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. 2014;19:251–72.
20. Sweller J. Cognitive Load During Problem Solving : Effects on Learning. 1988;285:257–85.
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(9):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 Oto-Rhino-Laryngology [Internet]. 2020;(0123456789). Available from: https://doi.org/10.1007/s00405-020-05858-3
23. Van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: Design principles and strategies. Med Educ. 2010;44(1):85–93.
24. Sethia R, Kerwin TF, Wiet GJ. Performance assessment for mastoidectomy: State of the Art Review Rishabh. Otolaryngol Head Neck Surg. 2017;156(1):61–9.
25. Al-Shahrestani F, Sørensen MS, Andersen SAW. Performance metrics in mastoidectomy training : a systematic review. Eur Arch Oto-Rhino-Laryngology [Internet]. 2019;276(3):657–64. Available from: http://dx.doi.org/10.1007/s00405-018-05265-9
26. Vilmann AS, Norsk D, Bo M, Svendsen S. Computerized feedback during colonoscopy training leads to improved performance : a randomized trial. Gastrointest Endosc [Internet]. 2018;88(5):869–76. Available from: https://doi.org/10.1016/j.gie.2018.07.008
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;(72):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;484–7.

29. Andersen SAW, Cayé-Thomasen P, Sørensen MS. Mastoidectomy performance assessment of virtual simulation training using final-product analysis. Laryngoscope. 2015;125(2):431–5.

30. Naismith LM, Cavalcanti RB. Validity of Cognitive Load Measures in Simulation-Based Training: A Systematic Review. Acad Med. 2015;90(11):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 [Internet]. 2016;73(1):45–50. Available from: http://dx.doi.org/10.1016/j.jsurg.2015.09.010

32. Leppink J. Data analysis in medical education research: a multilevel perspective. Perspect Med Educ [Internet]. 2015;4(1):14–24. Available from: http://link.springer.com/10.1007/s40037-015-0160-5

33. Zirkle M, Taplin MA, Anthony R, Dubrowski A. Objective Assessment of Temporal Bone Drilling Skills. Ann Otol Rhinol Laryngol. 2007;116(11):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 BT - Artificial Intelligence in Education. In: Isotani S, Millán E, Ogan A, Hastings P, McLaren B, Luckin R, editors. Cham: Springer International Publishing; 2019. p. 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. Otol Neurotol. 2015;152(6):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(3):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 [Internet]. 2019; Available from: https://doi.org/10.1016/j.joto.2019.12.003

38. Andersen SAW, Frendø M, Sørensen MS. Effects on cognitive load of tutoring in virtual reality simulation training. MedEdPublish. 2020;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. 2017;269–75.

TABLE & FIGURE LEGENDS

Table 1. Participant characteristics

	Intervention Structured summative feedback	**Control** No feedback
No. of participants	12	12
Age, mean (SD)	23 (1.4)	26 (9.9)
Sex, N		
Female	8	6
Male	4	6
Weekly computer usage excl. work (hours), mean (SD)	8.1 (6.8)	9.8 (5.3)
Self-reported computer skills (Likert scale 1-7), mean (SD)	5.1 (0.7)	4.3 (1.4)
Gaming frequency (Likert scale 1-5), mean (SD)	3.9 (1.1)	3.6 (1.4)

Table 2. Performance in retention testing.

	Mean final-product score, points	**Mean metrics-based score, %**	**Relative increase in cognitive load, %**
Intervention group	15.3 (95% CI [14.2 to 16.4])	81.5 (95% CI 77.5 to 85.4)	30.0 (95% CI 26.4 to 33.5)
Control group	14.4 (95% CI [13.3 to 15.4])	74.6 (95% CI 70.6 to 78.6)	20.0 (95% CI 16.5 to 23.6)
P-value	.23	.02	<.001

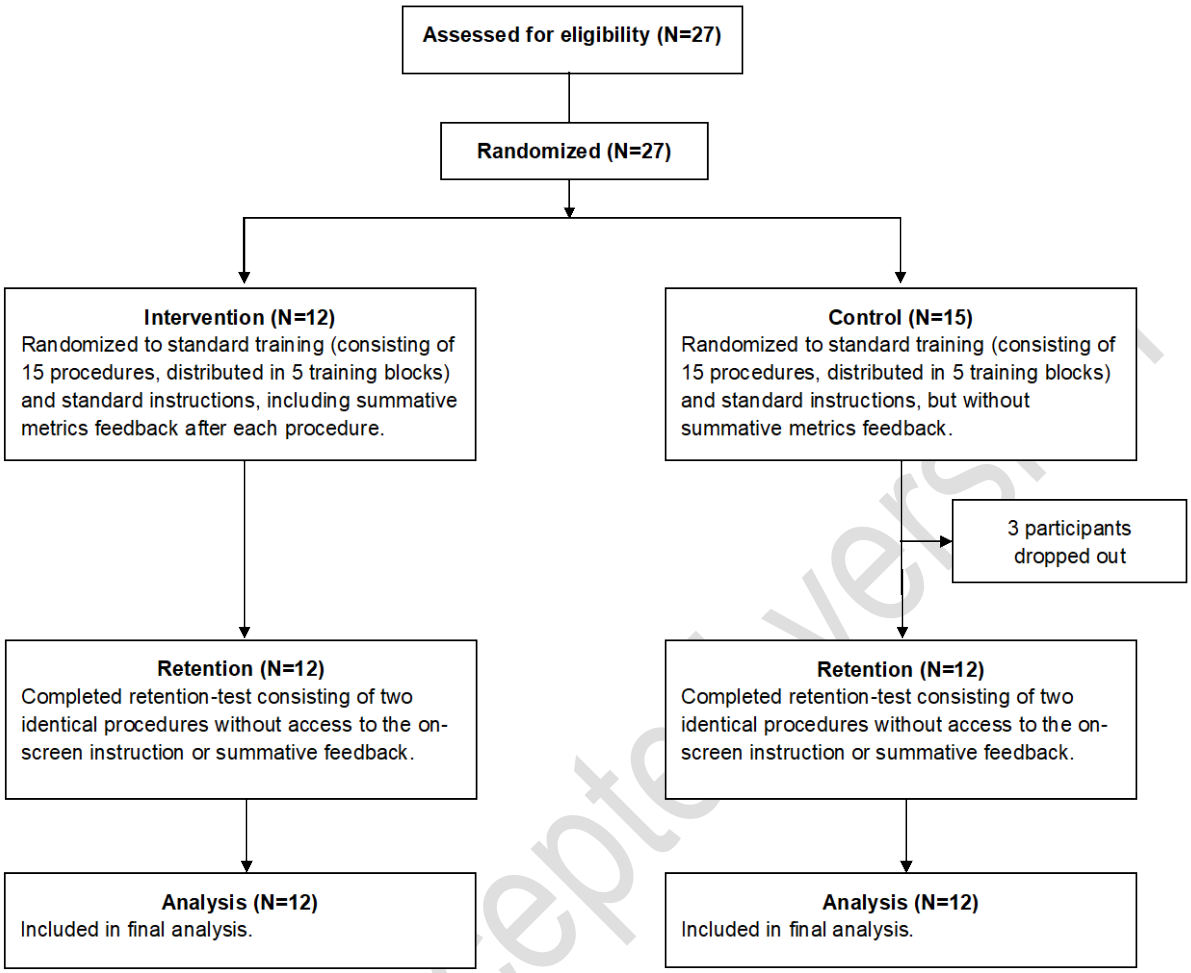


Figure 1. CONSORT flow diagram.

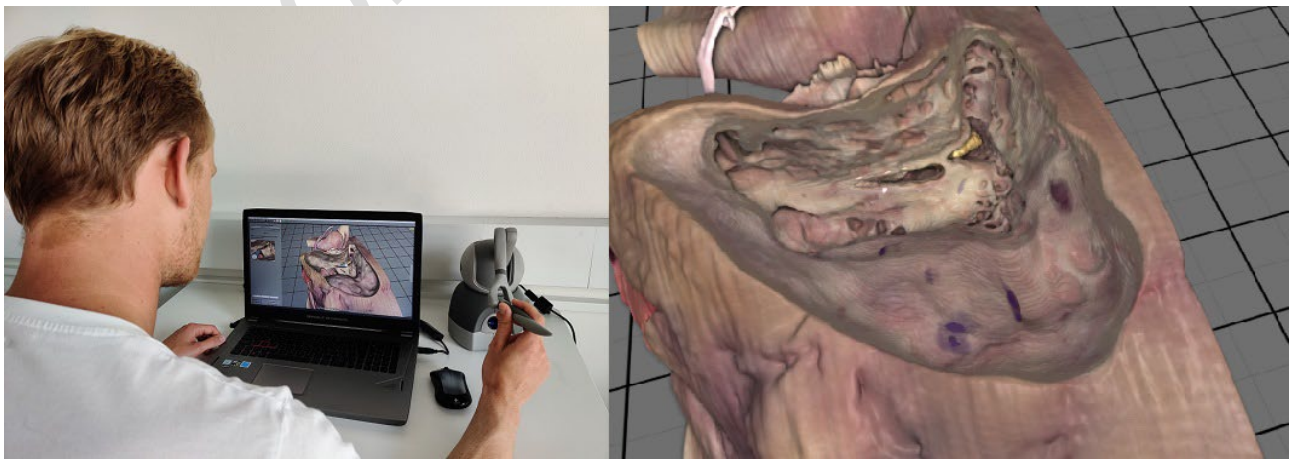


Figure 2. Simulation set-up (left) and an example of a mastoidectomy final-product after a training procedure (right).

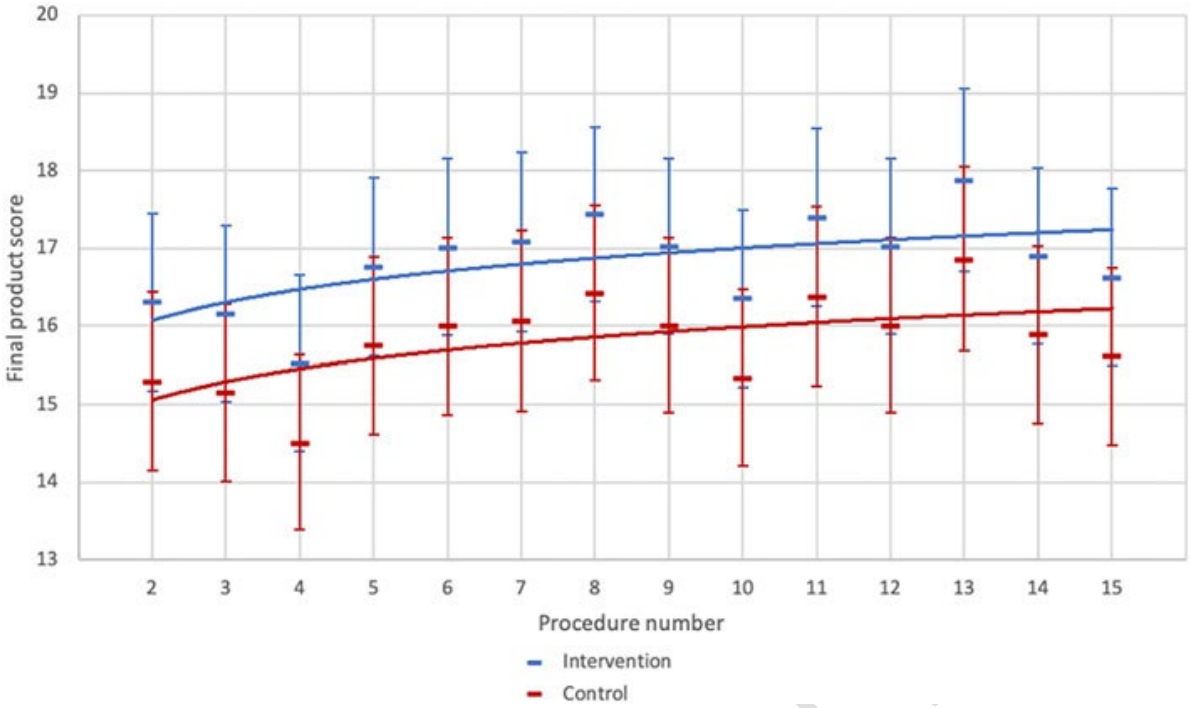


Figure 3. Final product score learning curves of training sessions (procedure 2–15). Means plot (estimated marginal means). Bars mark 95 % CI.

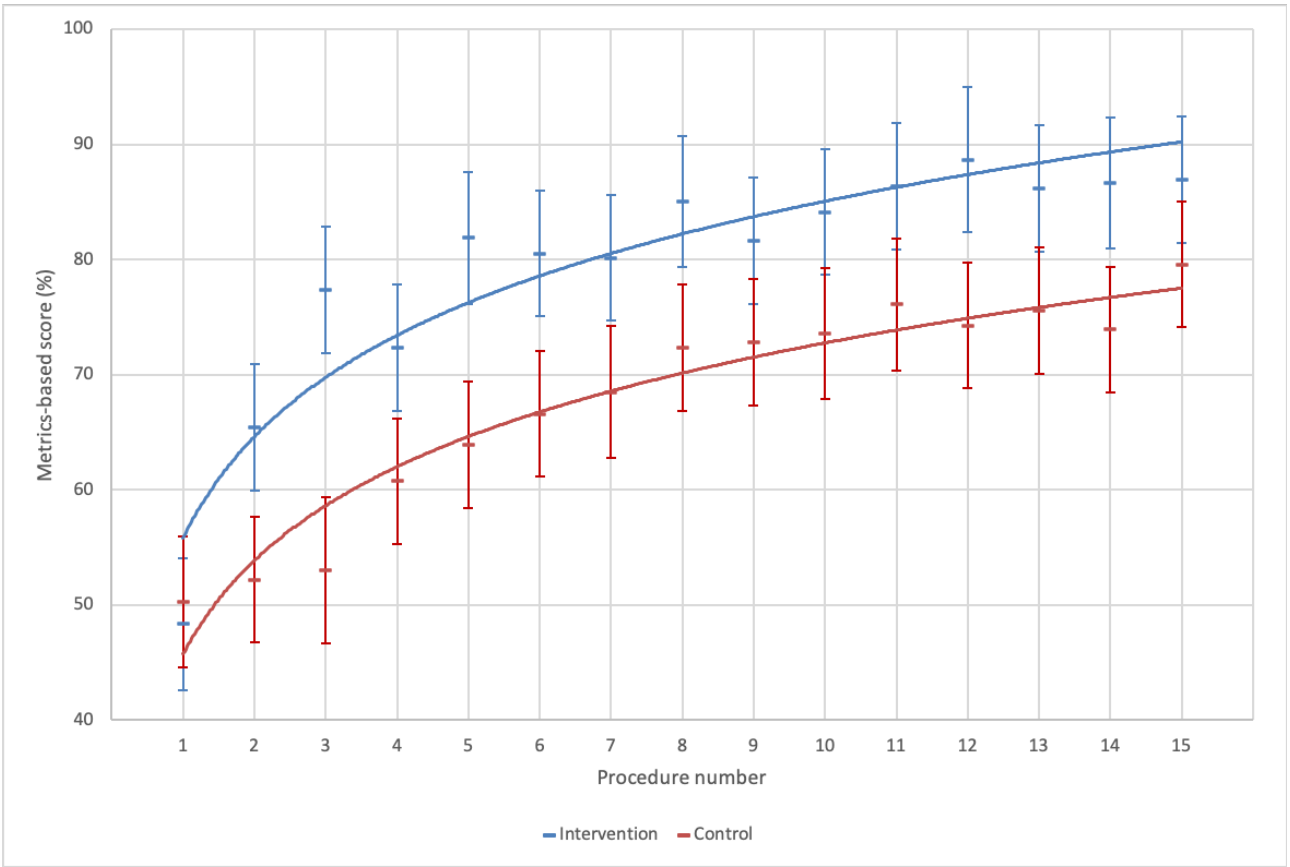


Figure 4. Metrics-based scores learning curves of training sessions (procedure 1–15). Means plot (estimated marginal means). Bars mark 95 % CI.