UNIVERSITY OF COPENHAGEN
FACULTY OF HEALTH AND MEDICAL SCIENCES

RESEARCH DOCTORAL DISSERTATION

Simulation-based training and assessment of mastoidectomy
—perspectives on the outside, inside, and in-between conditions of practice

Steven Arild Wuyts Andersen, MD, PhD

Copenhagen Hearing and Balance Center
Department of Otorhinolaryngology
Rigshospitalet, Copenhagen, Denmark.

The Faculty of Health and Medical Sciences at the University of Copenhagen has accepted this dissertation for public defence for the doctoral degree in Medicine. Copenhagen, 25 August 2023.

Bente Merete Stallknecht, Head of Faculty

The public defence will take place on 23 November 2023 at 15:00 in Auditorium 1 at Rigshospitalet, Blegdamsvej 9, Copenhagen, Denmark. Professor Steffen Heegaard, University of Copenhagen, has been appointed chair of the defence ceremony.

Members of the assessment committee:
Erik Driessen, professor, Maastricht University (1. opponent)
Michael Gaihede, professor, Aalborg University (2. opponent)
Preben Homøe, professor, University of Copenhagen (chair)

DOI: 10.5281/zenodo.8265247
ISBN: 978-87-974872-0-4

Preface and acknowledgements

Temporal bone surgery involves drilling of the lateral side of the skull to gain access to the middle ear for otologic procedures such as treatment of cholesteatoma and cochlear implantation. The use of virtual reality (VR) simulation for temporal bone surgical training offers a unique opportunity to improve the training of surgeons without the constraints such as the access to and availability of human cadaveric temporal bones for traditional dissection training.

During my PhD studies on VR simulation training of novices, which were kindly funded by the Oticon Foundation, I observed an early and inadequate performance plateau of repeated practice in the simulation. This fueled my interest in understanding the underlying mechanisms and to formulate and research different strategies to improve novice performance beyond the initial plateau and to further improve simulation-based training and explore it for clinical purposes. These efforts have ultimately led to this research doctoral thesis on perspectives on the conditions of practice in simulation-based training of mastoidectomy. However, my aim goes beyond temporal bone surgery: using the mastoidectomy procedure as exemplary, hopefully, this research can generalize to simulation of other complex surgical procedures and benefit trainees and patients alike.

This research doctoral thesis is based on studies conducted at the Department of Otorhinolaryngology—Head & Neck Surgery, Rigshospitalet; the Simulation Centre at Copenhagen Academy for Medical Education and Simulation (CAMES); and the Dept. of Otolaryngology, Nationwide Children's Hospital and the Ohio State University, Columbus, Ohio, USA. These studies were conducted in the period 2016–2020 as a postdoc during and in between my post-graduate clinical training towards becoming a specialist in otorhinolaryngology.

I would, first of all, like to recognize professor Mads Sølvsten Sørensen, who served as the main supervisor during my PhD studies, but who continues to be my mentor scientifically and clinically. His dedication to the Visible Ear Simulator project and visions for temporal bone simulation are extraordinary and serve as my greatest inspiration. I feel honored by his support throughout the process of my continued research and clinical training.

Next, I would like to express my profound gratitude and respect for the work of Peter Trier Mikkelsen, who programmed the Visible Ear Simulator in all its iterations, and who did the considerable programming that made this research possible. Further, I would like to acknowledge professor Lars Konge for his continuous support and also for providing me opportunities to develop further as a researcher as first a postdoc and later as an external lecturer at the Simulation Centre at CAMES. Finally, I need to thank my international mentor, Prof. Gregory J. Wiet, who hosted me for 15 months during my international postdoc fellowship funded by the Independent Research Fund Denmark. He welcomed me wholeheartedly into his team with Dr. Kimerly Powell and Bradley Hittle and I learned so much from their multidisciplinary approach to using clinical imaging datasets for temporal bone simulation.

I would also like to mention a number of other people who in various ways have contributed valuably to my research: Professor Per Cayé-Thomasen, who not only has strongly supported temporal bone simulation at the level of the Danish Society of Otorhinolaryngology, but who also together with Dr. Søren Foghsgaard has rated the temporal bone performances at the national temporal bone course for a number of years; the rest of the otology team at Rigshospitalet: Drs. Susan Jacobsen, Sune Bloch, Ramon Gordon Jensen, Martin Nue Møller, Martin Reznitsky, Christian Siim; professor Christian von Buchwald and head of department Mads Klokker and his secretary Berit Merete Sachmann for their support; from the Simulation Centre: Leizl Joy Nayahangan, Therese Møller-Andersen, Ditte Guldmann Kryger Rasmussen, Torben Schrøder, Anne Mette Mørcke, Sam Kondo Steffensen, Morten Bo Svendsen, Katrine Mortensen and other staff and researchers at CAMES; Dr. Yoon Soo Park, associate professor and medical educational statistical expert at Harvard Medical School; at the Ohio State University and

Nationwide Children's Hospital, Columbus, Ohio: Thomas Kerwin, Jason Keith, Grace Oswald, Varun Varadarajan, Aaron Moberly, Prashant Malhotra, Oliver Adunka—along with so many others from staff and administration who contributed to my research and helped with my research stay.

Finally, a special mention to my colleagues and fellow researchers Kristianna Mey, Eva Rye Rasmussen, Ebbe Thinggaard, Flemming Bjerrum, Ann-Sofia Skou Thomsen, Jacob Melchiors, and Martin Tolsgaard for our many academic discussions. Furthermore, my research students Mads Guldager, Joakim Grant Frederiksen, Lisette Hovgaard, Fahd Al-Sharestani, Josefine Buchwald, Karoline Arnesen to name a few who all deserve to be recognized for their amazing work and for developing me as a supervisor in their respective projects; and to Drs. Susanne Scott, Martin Frendø and Andreas Frithioff who trusted me enough to let me co-supervise their PhD projects and who have taught me a great deal in return.

None of this research would have been possible without the people who volunteered for participation: medical students from the Faculty of Health and Medical Sciences at the University of Copenhagen; my fellow colleagues and residents in otorhinolaryngology; and the expert surgeons from the Danish Otosurgical Society, the otology team in Columbus, and our collaborators in Sweden, Norway and across the USA, all who contributed with their performances and expertise to the project. I am grateful for their valuable contributions.

Finally, I want to express profound gratitude for my family and friends for their support throughout the process.

Steven Arild Wuyts Andersen, MD, PhD

April 2022

Summary

Mastoidectomy is a complex surgical procedure that involves drilling of the temporal bone of the skull to gain access to the middle ear and surrounding structures. Hands-on training through traditional cadaver dissection is increasingly difficult to provide given the diminishing number of donated human temporal bones and the high number of trainees. Modern surgical education also requires evidence-based approaches to skills training and competency assessment. Altogether, this has propelled the development of virtual reality (VR) simulation for surgical skills training including in temporal bone surgery.

It is well-established that VR temporal bone surgical simulation is highly useful for the training of novices. However, the performance of trainees seems to plateau early and at an insufficient level under self-directed training conditions. We therefore need to better understand what elements make self-directed simulation-based training work in order to design efficient training programs and create high-quality learning experiences. In this thesis, different learning conditions based on contemporary medical educational frameworks were studied because what we do outside, inside and in-between simulation matters.

The principle of mastery learning requires repeated practice until the level of proficiency. This motived us to develop a metrics-based performance assessment for automated assessment and to define pass/fail standards of proficiency using the expert performance framework. Nevertheless, the metrics-based score failed to capture key indicators of a safe performance compared with the established but manual and time-consuming final-product assessment. Next, generalizability theory was used to explore the reliability of the established final-product assessment under different training conditions: Contextual variables such as simulation modality and fidelity considerably affected reliability, cautioning the use of established assessment tools under conditions and contexts different from their original setting without specific considerations on reliability.

Distribution of practice over time is in the motor skills literature recognized to be superior to massed practice for skills acquisition and we have previously corroborated this for VR simulation training of temporal bone surgery. In this thesis, the effect of supplemental distributed VR simulation training on transfer of skills to the cadaveric dissection simulation training modality was therefore investigated: five training blocks of three simulated procedures improved subsequent cadaveric dissection performance by 25 %.

Cadaveric dissection represents a more complex learning environment and learning task than the VR simulation condition and further we found that the learning conditions of cadaveric dissection induces a significantly higher cognitive load in trainees. According to cognitive load theory, a cognitive load that exceeds the capacity of the learner can negatively affect learning. Therefore, the effects of repeated practice on cognitive load were explored and in contrast to massed practice, distributed practice significantly reduced cognitive load.

Altogether, these findings have the implication for the temporal bone surgical curriculum that training should be organized with structured and distributed VR simulation first to optimize the subsequent use of the limited and costly human temporal bones for dissection training after basic skills have been acquired.

VR simulation training allows distributed and repeated practice at the individual trainee's convenience but without instructor presence, other learning supports for feedback are needed to ensure a successful learning experience. The concept of directed, self-regulated learning emphasizes the importance of providing the learner with direction and guidance to support and scaffold training.

Simulator-integrated tutoring by green-lighting is an innovative approach to dealing with this challenge. However, the use of simulator-integrated tutoring in VR temporal bone surgical simulation was previously found to lead to tutoring over-reliance, causing a poor performance once tutoring was discontinued. Distributed practice was found to have a moderately protective effect against this phenomenon and we therefore hypothesized that intermittent simulator-integrated tutoring would be a better strategy. Consequently, the effect of

intermittent tutoring in a distributed training program was studied and it was found that tutoring increased performance while active, but resulted in an inferior performance in subsequent non-tutored sessions compared with a never-tutored reference cohort. This tutor over-reliance degrades motor skills learning and concurrent feedback through simulator-integrated tutoring should be reconsidered.

We next explored increasing fidelity of the VR simulation to better bridge the gap between simulation-based training and real-life conditions where an operating microscope is used to magnify the surgical field and to enable visualization of minute visual cues such as the vasculature of underlying anatomical structures. However, the improvement in resemblance and functional task alignment of introducing the eyepiece from a digital operating microscope did not benefit learners: compared with learners who were randomized for the conventional screen-based VR simulation condition, the learners in the "ultra-high fidelity" condition performed significantly poorer and their CL was higher. Consequently, improving instructional design and other learning supports should be considered over increasing realism of simulation.

Finally, we used an automated pipeline for segmentation of clinical CBCT imaging of the temporal bone to create patient-specific VR simulation that can be used for surgical rehearsal and planning ahead of actual surgery. Clinicians rated the patient-specific simulation highly, found that it contributed to a better understanding of the patient's anatomy, and perceived it to be of benefit to training of surgeons at both the resident and fellow level. Nonetheless, a major limitation was the quality of the clinical scans, which were often limited by field-of-view and poor scan quality due to motion artefacts.

Altogether, the work presented in this thesis provides insights into the outside, inside and in-between conditions of VR simulation training in temporal bone surgery—with broader implications for simulation-based surgical skills training in general. The optimal VR simulation training program consists of structured and distributed practice, supporting directed, self-regulated learning. The role of addressing the cognitive process, motivating the trainee, and providing proper direction and feedback cannot be stressed enough.

Future research directions include developing an adaptive training program that tailors feedback and case difficulty based on valid and reliable automated assessment, better integration into the clinical training curriculum, and advancing simulation for training beyond the novice level so it becomes a useful tool even for patient-specific rehearsal and surgical planning and navigation.

Resumé

Mastoidektomi er en kompleks kirurgisk procedure der involverer udboring af kraniets tindingeben for at skabe adgang til mellemøret og omkringliggende strukturer. Hands-on træning gennem traditionel kadaver dissektion er tiltagende vanskeligt at sikre grundet det aftagende antal af donerede tindingeben og det høje antal uddannelsessøgende. Ydermere stilles der i stigende grad krav om evidensbaseret kirurgisk træning og kompetencevurdering. Dette har drevet udviklingen af virtual reality (VR) simulation til kirurgisk træning herunder også i tindingebenskirurgi.

Det er veletableret at VR-simulation er særdeles velegnet til tindingebenskirurgisk træning af novicer. Imidlertid når de uddannelsessøgende hurtigt et plateau på et utilfredsstillende lavt niveau når de træner på egen hånd i simulatoren. Der er derfor brug for at vi får en bedre forståelse for hvilke elementer af træningsprocessen der kan forbedres med det formål at designe effektive træningsprogrammer og understøtte bedre læring. I denne afhandling undersøges forskellige strategier baseret på aktuel medicinsk uddannelsesteori for at opnå indsigt i effekterne på den kirurgiske præstation og læringsprocessen fordi hvad vi gør uden for, inde i og ind imellem simulationstræningen er af betydning.

Princippet om mestringslæring inkluderer gentagen træning indtil et foruddefineret færdighedsniveau. Dette tilskyndede os til at udvikle en metrik-baseret præstationsscore med henblik på automatiseret bedømmelse og til at definere beståelsesgrænser ud fra eksperters præstation. Desværre kunne denne metrik-baserede score ikke indfange nøgleindikatorer omkring en sikker præstation sammenholdt med den etablerede slutproduktbedømmelse, der til gengæld kræver manuel vurdering. Generaliserbarhedsteori blev brugt til at undersøge pålideligheden af slutproduktbedømmelse under forskellige træningskonditioner: kontekstvariable havde betydelig effekt på pålideligheden, hvilket understreger at etablerede kompetencevurderingsredskaber ikke bør anvendes under andre betingelser og i andre kontekster end de oprindeligt var udviklet til uden specifikke overvejelser omkring ændring af pålideligheden.

Fordeling af træning over tid er i litteraturen omkring motorisk færdighedstræning anerkendt som overlegen til kondenseret træning. Dette har vi tidligere vist også gør sig gældende for VR-simulationstræning af tindingebenskirurgi. Derfor ville vi yderligere undersøge om effekterne af distribueret træning ville resultere i en bedre dissektionspræstation – såkaldt transfer: fem træningsblokke á tre simulerede procedurer forbedrede den efterfølgende dissektionspræstation med 25 %.

Dissektionen udgør et meget komplekst læringsmiljø og en mere kompleks læringsopgave end VR-simulation og dissektionstræningen blevet fundet at medføre et betydeligt højere cognitive load hos de uddannelsessøgende. Ifølge cognitive load teori kan et cognitive load der overgår den mentale arbejdskapacitet forringe læring. Effekterne af distribueret og gentagen træning på cognitive load blev derfor undersøgt og vi fandt at dette signifikant reducerede cognitive load sammenholdt med kondenseret træning. Sammenfattet bør det kirurgiske uddannelsesprogram baseres på struktureret træning distribueret over tid for at optimere udbyttet af den begrænsede og dyre dissektionstræning.

VR-simulationstræning muliggør distribution af træningen tilpasset den enkelte uddannelsessøgendes behov og skema, men uden tilstedeværelse af instruktører er der behov for andre støtteredskaber til feedback for at sikre effektiv læring. Ifølge konceptet "målrettet, selvreguleret læring" er det altafgørende at sikre et klart mål og effektive instruktioner ved egen-læring. En simulatorintegreret tutorfunktion med visuel fremhævning af det enkelte trin i proceduren er en innovativ tilgang til denne udfordring. Problemet er dog at dette kan medføre afhængighed af tutorfunktionen, hvilket resulterer i en ringere præstation når tutor-funktionen efterfølgende gøres utilgængelig. Distribueret træning beskytter moderat mod denne negative effekt og vi opstillede derfor en hypotese om at intermitterende brug af tutorfunktionen ville være optimal for læringen. Derfor blev en kohorte af uddannelsessøgende som intermitterende gjorde

brug af tutorfunktionen sammenlignet med en referencekohorte der ikke fik adgang til tutorfunktionen. Tutorfunktionen øgede præstationen mens den var aktiv men til gengæld resulterede i ringere præstationer i efterfølgende øvelser, hvor tutorfunktionen var utilgængelig. Dette indikerer at feedback i form af den simulatorintegrerede tutorfunktion forringer indlæringen af motoriske færdigheder og bør derfor gentænkes.

Dernæst undersøgte vi effekten af at øge realismen af VR-simulationen for bedre at bygge bro til virkelighedens forhold, hvor man anvender et operationsmikroskop til at forstørre det kirurgiske felt og for at kunne se afgørende detaljer som fx karstrukturen over anatomiske landemærker. Desværre var tilføjelsen af okularet fra et digitalt operationsmikroskop mhp. at øge realismen ikke en fordel for læringen: sammenlignet med deltagere der blev randomiseret til konventionel, skærmbaseret VR-simulation, var præstationen signifikant dårligere og cognitive load højere blandt deltagerne i denne "ultra-high fidelity" simulation med okularet. Derfor bør man overveje at forbedre læringen ved brug af andre metoder fremfor at øge realismen af simulationen.

Endeligt anvendte vi en automatiseret proces til segmentering af kliniske CBCT-scanninger af tindingebenet for at skabe patient-specifik VR-simulation der kan bruges til kirurgisk indøvning og planlægning forud for det faktiske indgreb. Klinikere var positivt stemt overfor den patient-specifikke simulation, fandt at den øgede deres forståelse for patientens anatomi og opfattede simulationen som gavnlig for træning af uddannelsessøgende kirurger. Ikke desto mindre udgjorde kvaliteten af de kliniske scanninger en begrænsning pga. scanningsfeltet og bevægeartefakter.

Samlet præsenterer denne afhandlings arbejder indsigt i betydningen af hvad vi gør uden for, inde i og ind imellem VR-simulationsbaseret træning indenfor tindingebenskirurgi med bredere perspektiver for tilrettelæggelsen af simulationsbaseret træning af kirurgiske færdigheder generelt. Det optimale VR-simulationstræningsprogram baserer sig på princippet om understøttet og målrettet selvreguleret læring med mulighed for træning distribueret over tid. Vigtigheden af grundige overvejelser om at understøtte kognitive processer, at motivere den uddannelsessøgende, og sørge for instruktion og feedback i træningstilrettelæggelsen kan ikke understreges nok.

Fremtidige forskningsperspektiver inkluderer udviklingen af et adaptivt træningsprogram, som tilpasser feedback og sværhedsgrad baseret på pålidelig automatisk kompetencevurdering, en bedre integration af simulationsbaseret træning i det kliniske uddannelsesprogram, og at videreudvikle simulation så det også er værdifuldt for den aktive kliniker til patientspecifik træning, kirurgisk planlægning og navigation.

Table of Contents

1. Introduction
    1.1 Temporal bone surgery and training 1
    1.2 Directed, self-regulated learning 1
    1.3 Performance assessment 2
    1.4 Learning curves and skills transfer 3
    1.5 Cognitive load 5
    1.6 Clinical imaging of the temporal bone and segmentation 6
    1.7 Aims and objectives 7
2. Simulation-based assessment
    2.1 Background 8
    2.2 Using the expert framework for developing a metrics-based score and standard setting of performance 9
    2.3 Exploring the reliability of simulation-based assessment 10
3. Training organization
    3.1 Background 12
    3.2 The effects of distributed VR simulation training on transfer 12
    3.3 The effects of VR simulation training on CL 13
4. Simulator-integrated tutoring
    4.1 Background 14
    4.2 The effects of tutoring on performance 15
5. Fidelity of simulation
    5.1 Background 16
    5.2 The effect of ultra-high-fidelity VR simulation on performance and CL 17
6. Patient-specific simulation
    6.1 Background 18
    6.2 The potential of patient-specific simulation 19
7. Discussion
    7.1 Summary of main findings and implications 20
    7.2 Comparison with current literature 22
    7.3 Limitations 25
    7.4 Perspectives and future directions 28
8. Conclusion 29
9. References 29

Abbreviations

CBCT	Cone-Beam Computerized Tomography
CI	Cochlear Implant/Implantation
CL	Cognitive Load
CLT	Cognitive Load Theory
CT	Computerized Tomography
DSRL	Directed, Self-Regulated Learning
FPA	Final Product Assessment/Analysis
FPS	Final Product Score
G	Generalizability
GRS	Global Rating Scale
MBS	Metrics-Based Score
MRI	Magnetic Resonance Imaging
RCT	Randomized, Clinical Trial
SBT	Simulation-Based Training
TBC	Task-Based Checklist
VR	Virtual Reality

List of included papers

This thesis is based on the following papers, none of which have previously formed part of a thesis or degree:

I. Andersen SA, Mikkelsen PT, Sørensen MS. Expert sampling of VR simulator metrics for automated assessment of mastoidectomy performance. *Laryngoscope*. 2019; 129(9):2170–2177.

II. Andersen SA, Park YS, Sørensen MS, Konge L. Reliable assessment of surgical technical skills is dependent on context: an exploration of different variables using Generalizability theory. *Acad Med*. 2020; 95(12):1929–1936.

III. Andersen SA, Fogshgaard S, Cayé-Thomasen P, Sørensen MS. The effect of a distributed virtual reality simulation training program on dissection mastoidectomy performance. *Otol Neurotol*. 2018; 39(10):1277–1284.

IV. Andersen SA, 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(1):45–50.

V. Andersen SA, Mikkelsen PT, Konge L, Cayé-Thomasen P, Sørensen MS. Cognitive load in distributed and massed practice in virtual reality mastoidectomy simulation. *Laryngoscope*. 2016; 126(2):E74–9.

VI. Andersen SA, Mikkelsen PT, Sørensen MS. The Effect of simulator-integrated tutoring for guidance in virtual reality simulation training. *Simul Healthc*. 2020; 15(3):147–153.

VII. Frithioff A, Frendø M, Mikkelsen PT, Sørensen MS, Andersen SA. Ultra-high-fidelity virtual reality mastoidectomy simulation training—a randomized, controlled trial. *Eur Arch Otorhinolaryngol*. 2020;277(5):1335–1341.

VIII. Andersen SA, Varadarajan VV, Moberly AC, Hittle B, Powell KA, Wiet GJ. Patient-specific Virtual Temporal Bone Simulation Based on Clinical Cone-beam Computed Tomography. *Laryngoscope*. 2021;131(8):1855–1862.

1. Introduction

1.1 Temporal bone surgery and training

Temporal bone surgery comprises surgical procedures related to the temporal bone of the skull with mastoidectomy being the principal procedure. Mastoidectomy involves the gradual drilling of the temporal bone to gain access to the middle ear and anatomical structures localized within the temporal bone. The procedure can be modified depending on the purpose of surgery and be combined with for example posterior tympanotomy to provide access to the round window for cochlear implantation (CI).

Temporal bone surgery requires complex psychomotor skills and these have traditionally been acquired through dissection of human cadaveric temporal bones and supervised surgery on patients in the operating room, i.e., surgical apprenticeship.[1] Cadaveric dissection can be considered a high-fidelity simulation model and is still considered the gold-standard training modality in temporal bone surgery.[2] Human cadavers for training are an increasingly scarce resource and only few but the largest training institutions can offer trainees in-house access to open dissection facilities and unlimited human temporal bones.[3] Therefore, most training departments instead send trainees on brief temporal bone courses that most often offer only cadaveric dissection training.[3]

Several alternative training models have been proposed including artificial models of plaster or plastic, technology-enhanced simulation such as virtual reality (VR) simulation, or hybrid models that augment physical models with virtual elements. The development of VR simulators were driven by the advances in computer graphics and the earliest models were applications based on polygon surface models typically used in gaming.[4,5] In contrast to other surgical approaches such as laparoscopic surgery and endoscopy, surface models were quickly abandoned in VR temporal bone simulation because drilling requires a volumetric approach to better simulate the gradual removal of bone.[6–9] These volumetric models are typically based on CT imaging data and amongst these, the still active projects in the field are, respectively, the Voxel-Man Tempo (Hamburg, Germany), the CardinalSim (Stanford University, California, USA), the University of Melbourne (Melbourne, Australia), and the Ohio State University (Columbus, Ohio, USA) temporal bone simulators. In contrast, the primary model in the Visible Ear Simulator is based on high-resolution digital cryosection images,[10] providing highly realistic graphic detail. This has recently been supplemented by several other temporal bone models based on a combination of cone-beam CT (CBCT) and micro-slicing of cadaveric temporal bone specimens.[11,12] The Visible Ear Simulator is offered as academic freeware for download[13] and has been detailed in several of our publications.[14,15]

Evidence for the effect of VR simulation for temporal bone surgical training of novices has accumulated in the published literature across the different VR temporal bone surgical platforms since 2007.[15] However, a main challenge is how to overcome the early plateau in the performance of novices in VR simulation-based training (SBT) in temporal bone surgery.[16,17] Reliable assessment of performance is a prerequisite in order to investigate the effect of different learning conditions. Finally, little is known on the value of simulation for more experienced temporal bone surgeons.

We have a unique opportunity in VR simulation of temporal bone surgery because of strong academic projects focused on research-based simulation. This allows us to go beyond what can typically be studied using commercially available VR simulators because we can easily modify every aspect of the simulation including learning supports and metrics measured. Hopefully, the achievements in VR simulation of temporal bone surgery can in turn be generalized to other procedures and inspire development towards clinical use of simulation for patient-specific simulation.

1.2 Directed, self-regulated learning

SBT is at the core hands-on and learner centred. One of the potential advantages of SBT of surgical technical skills, and especially VR simulation training, is that the trainee can practice autonomously with no need for the presence of a human instructor, i.e., "unsupervised".[18,19] This substantially reduces the

costs of SBT and allows the trainee to practice according to individual needs and schedule. If clear educational goals are defined, the trainee can work freely within the learning environment and self-regulate their learning to achieve these goals. An educational goal can for example be training to the level of proficiency[20] after which the learner needs to further refine their skills using more advanced training models or advance to clinical training.

Proficiency-based training is frequently used as the terminology for applying the principles of mastery learning in the context of SBT of surgical technical skills. The concept is basically that the learner practices until they can demonstrate a consistent performance at a pre-defined level. Such mastery learning seem to positively affect outcomes of simulation-based medical education.[21–23] However, the amount of practice needed to achieve proficiency is highly individual,[24] which is impractical if instructors need to be physically present during every practice session. Further, conventional instructor-led SBT may result in inferior self-regulated learning compared with a self-directed approach with supports for self-regulated learning.[25]

Self-regulated learning can be unguided or minimally guided (discovery approaches) or can be supported through the use of learning supports for scaffolding the learning experience. The concept of directed, self-regulated learning (DSRL) emphasises that the learner is provided with direction and guidance in order to achieve efficient self-directed learning.[26,27] A clear advantage of DSRL is cost effectiveness because of the reduced need for instructor presence.[28]

A major challenge of "unsupervised" and self-regulated learning is that without corrective feedback there is a risk that the trainee learns the wrong things, do them in the wrong way, or makes incorrect judgments.[29] Consequently, DSRL stresses the deliberate design of training with mechanisms to ensure a successful learning experience. Designing, implementing, investigating, and understanding different learning supports for optimal directed, self-regulated learning in the context of VR SBT is therefore highly relevant.

1.3 Performance assessment

Surgical skills have traditionally been documented through quantitative measures such as logging the number of procedures and hours of training, by written or oral tests that mainly reflect knowledge domains, and/or by direct or video-based observation of the operating room performance with feedback from the surgical supervisor (expert opinion).[1] The paradigm of competency-based medical education,[30] the introduction of the objective, structured assessment of surgical technical skills tool (OSATS),[31] and the use of simulation-based assessment,[32] have in the recent decades transformed surgical skills training and assessment profoundly. Importantly, evidence of validity and reliability of simulation-based assessment needs to be accumulated from a number of sources based on modern validity frameworks such as Messick's.[33] Valid and reliable assessment is fundamental for evaluating the effect of educational interventions and in any effort to systematically advance surgical education and training.

In temporal bone surgery, different approaches to assess mastoidectomy performance have been suggested. These include global rating scales of performance (GRS), structured task-based check lists (TBC), and final-product assessment (FPA).[34] However, there is conflicting evidence on the correlation between performance scores on different instruments,[34–36] most likely reflecting that process, technical skill, and end-result are separate domains. In contrast to global rating and checklists, which require somewhat lengthy direct or video-based observation of performance, the FPA considers only the end result of the temporal bone drilling.[37] Consequently, FPA only indirectly reflects process and motor skills performance, which is a clear limitation. Nonetheless, FPA is the most commonly reported outcome measure in the temporal bone surgical training literature[38] because it allows assessment to be performed at a later point in time. This makes standardized assessment of a large number of performances by several raters feasible. We have therefore consistently used final-product analysis using the Modified Welling Scale (Table 1) as a performance outcome throughout this thesis.

Although validity evidence has accumulated for all three types of instruments (GRS, TBC, and FPA) for mastoidectomy performance assessment in different training contexts (intra-operatively, cadaver dissection, and VR simulation) there is still little validity evidence on response process and consequences of assessment including standard setting for proficiency-based training.[38]

Technology-enhanced simulation including VR simulation also provides new opportunities for performance assessment—the use of computer-gathered metrics for automated and truly objective assessment. Metrics are basically any parameter that can be registered during a performance such as path length, force used, energy applied, collisions, errors etc.[39] Many of these metrics cannot easily be measured or tracked in real-life, complicating validation. Using this information for meaningful feedback and assessment is perhaps even more difficult and complexity increases further when considering derivatives of metrics for example path length per time, and aggregated scores of incommensurable metrics. Many commercial simulators incorporate such complex metrics for automated assessment. However, the same scientific rigor must be cautioned for metrics-based assessment as for more traditional assessment and validity evidence must be meticulously collected.

Table 1. Modified Welling Scale for final product analysis.

0 = incomplete/inadequate dissection 1 = complete/adequate dissection		
Mastoidectomy margins defined at		
1. Temporal line	0	1
2. Posterior canal wall	0	1
3. Sigmoid sinus	0	1
Antrum mastoideum		
4. Antrum entered	0	1
5. Lateral semicircular canal exposed	0	1
6. Lateral semicircular canal intact	0	1
Sigmoid sinus		
7. Exposed, no overhang	0	1
8. No cells remain	0	1
9. No holes	0	1
Sinodural angle		
10. Sharp	0	1
11. No cells remain	0	1
Tegmen mastoideum/tympani		
12. Attic/tegmen tympany exposed	0	1
13. Ossicles intact (untouched)	0	1
14. Tegmen mastoideum exposed	0	1
15. No cells remain	0	1
16. No holes	0	1
Mastoid tip		
17. Digastric ridge exposed	0	1
18. Digastric ridge followed towards stylomastoid foramen	0	1
19. No cells remain	0	1
External auditory canal		
20. Thinning of the posterior canal wall	0	1
21. No cells remain	0	1
22. No holes	0	1
Facial nerve		
23. Facial nerve identified (vertical part)	0	1
24. No exposed nerve sheath	0	1
25. Tympanic chorda exposed	0	1
Posterior tympanotomy		
26. Facial recess completely exposed	0	1

1.4 Learning curves and skills transfer

There are many definitions on the construct of "learning" but common features include that it is a "*transformative process*" resulting in "*a change in knowledge and/or behaviour*" that is "*persistent*". Further, the term learning can refer both to the product (end result), the process, or the function. Unsurprisingly, defining outcomes of learning experiences depends on the specific context and the purpose. In surgical skills training, learning outcomes often include immediate performance (i.e., here-and-now performance during training), learning curves (i.e., how performance changes with cumulative experience), retention of performance (i.e., preservation of performance after a period of non-rehearsal), and transfer (i.e., how performance in one context affects performance in another context). This is of course a gross oversimplification but an in-depth discussion of all these concepts is beyond the scope of this thesis. Nonetheless, this discourse highlights that there is a qualitative difference between performance and learning and that any systematic attempt to improve the outcome of training programs requires insights into the processes and mechanisms of learning. Ultimately, this cannot be achieved using a single outcome but rather requires triangulation from multiple sources.[40] In this thesis, this includes different measures of performance to investigate learning curves and transfer, and cognitive load (CL) to explore the cognitive processes involved in learning.

Learning curves describe the relationship

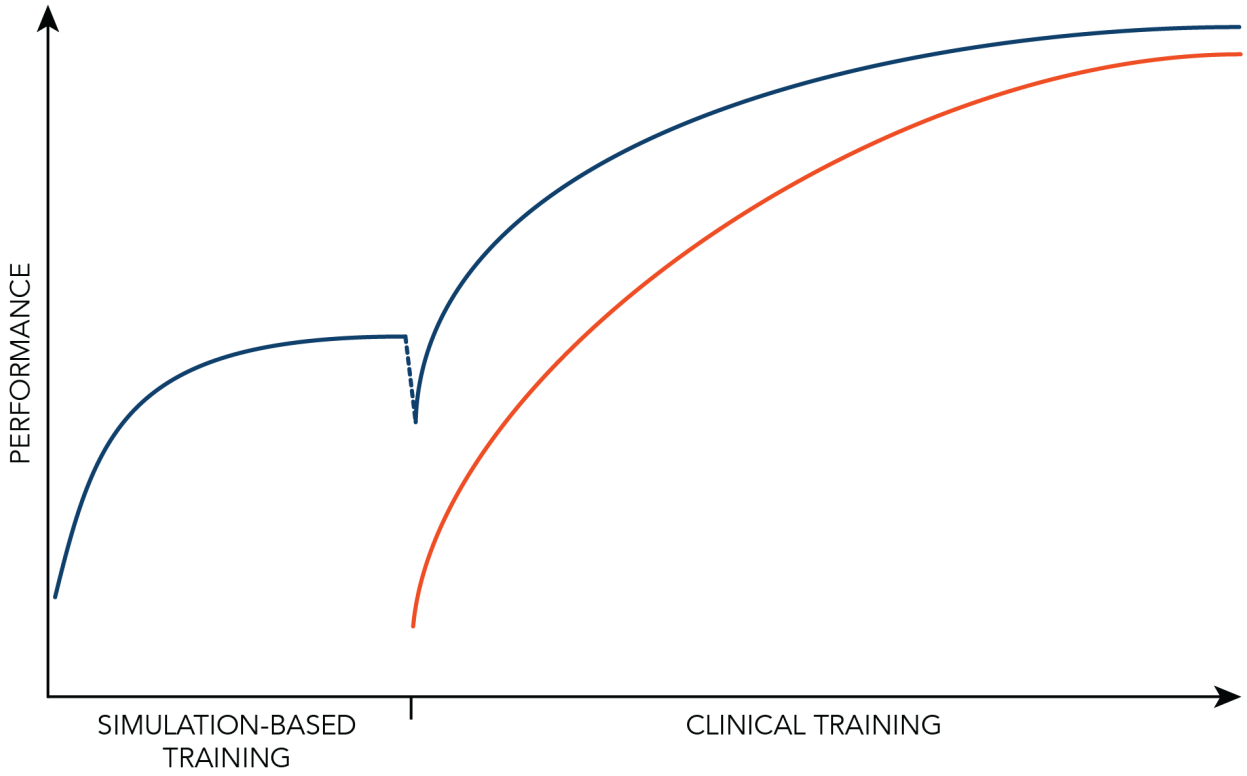

Figure 1. A learning curve model for simulation-based medical education. Modified after Konge et al.[43]

between experience and outcomes of learning.[41] The outcome of learning will typically increase with cumulative experience (amount of practice) following a negatively accelerated learning curve as described in the motor skills literature.[42] This configuration has some key features that include *slope* (rate of learning) at any given point and an *inflection* point where learning becomes more effortful, ultimately resulting in a *plateau* phase.[41] The configuration of the learning curve will be dependent on the learner including learner characteristics and previous experience, the learning task, contextual factors including the learning environment and instructional design, and the measured learning outcome. The learning curve of the individual learner rarely fits with the average for a group of learners under similar conditions and the averaged learning curve therefore represents an idealised model.[41] SBT of surgical technical skills is typically introduced before clinical training of the skill with the purpose of offsetting the learning curve of the subsequent clinical training (Figure 1).

There are many features of SBT that are associated with positive effects on learning including range of difficulty in training cases, repeated and distributed practice, cognitive interactivity, support for self-regulated learning, and feedback.[44] However, how these features work, for which groups of learners, and in which contexts remain under-investigated.[44] The learning curve configuration can provide insights useful in understanding how instructional design interventions affect the learning process and how SBT can be optimized.

In medical education, replicating knowledge (memory tasks) is typically less useful than applying established knowledge and skills to new problems, situations and domains—so-called transfer. The ultimate goal of SBT is that the skills learned in simulation results in a better performance in the clinical setting, resulting in better patient outcomes and safety. Minimizing the loss of performance due to transfer processes is therefore highly desirable.

At its core, transfer requires complex cognitive processes that are not well understood and are particularly difficult if the transfer task involves dissimilar problems, situations and/or domains.[45] A

cohesive understanding of transfer and the role of factors of the individual such as motivation and cognitive skills, contextual factors such as feedback, and task factors including instructional strategies is needed.[46] The efficiency of transfer could also depend on curriculum factors such as discontinuity between achieving competency in SBT and continuing to refine those skills in the clinical environment.

In temporal bone surgery in Denmark, transfer from the simulation environment to the OR is not a feasible research outcome measure because very few otorhinolaryngology residents progress to supervised temporal bone surgery. Since cadaveric dissection represents a more complex learning environment than VR simulation, we used this as an intermediate transfer outcome (reflecting *context transfer*[47]).

1.5 Cognitive load

Cognitive load theory (CLT) is a leading theory on educational instructional design based on knowledge of the human cognitive architecture and processes such as how we learn, think and solve problems.[48] Briefly, the underlying premise of CLT is that working memory is limited and can only process few elements simultaneously within a short temporal frame. The organized cognitive schemata stored in long-term memory can be used to offset these working memory limitations.[49] In the lens of CLT, expertise is the result of effective cognitive schemata that have been refined and automated through repeated and deliberate practice.

In contrast, novel information that does not fit within existing schemata require crude processing in working memory. If the task exceeds the capacity of the learner, cognitive overload can ensue and actually hinder the formation of efficient schemata.[48] This results in a challenge especially when dealing with novices as is often the case for example in SBT of surgical skills.

Cognitive Load (CL) is mostly considered to consist of three components (Figure 2): the intrinsic CL of the task, which is dependent on the task itself (the element interactivity) and the expertise of the learner (i.e., how well-developed and well-employed are the relevant mental schemata); germane CL that

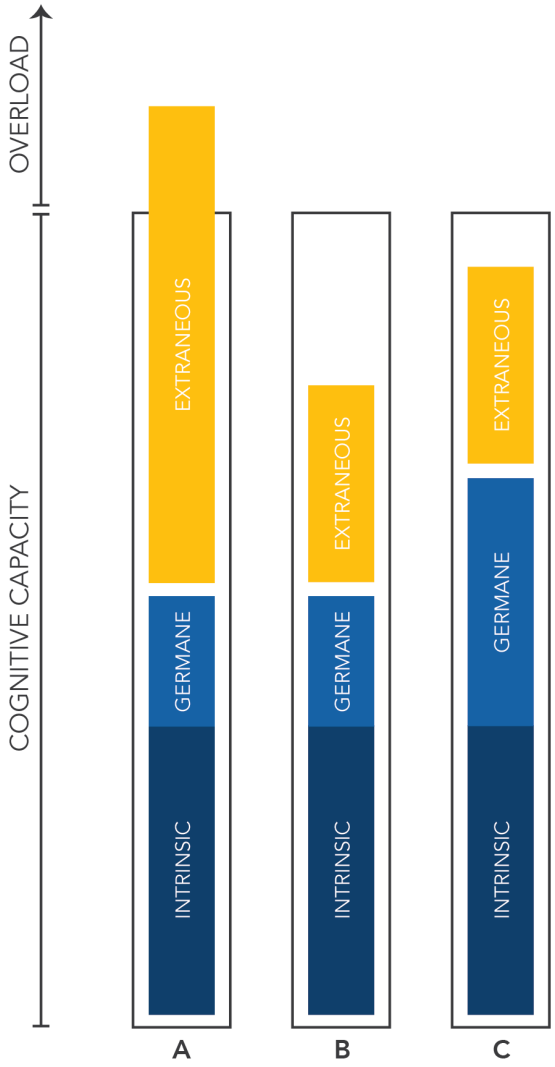

Figure 2. Cognitive load model after *van Merriënboer and Sweller*.[49] A. Cognitive load due to excessive extraneous load; B. Cognitive capacity not fully utilized; C. Increasing germane load allows full utilization of available cognitive capacity.

is used for dealing with the learning task such as constructing mental schemata (i.e., processes leading to actual learning); and extraneous CL that is unnecessary for learning and might result from weak problem solving methods.[49] The effects of the intrinsic and extraneous load components are largely regarded as additive, whereas the effect of germane load seems to be more like an interaction effect.[50] Altogether, these considerations lead to a number of proposed instructional design principles and strategies to reduce extraneous load, manage intrinsic load and optimise germane load in order to optimize

the learning experience.[51]

CL is "*a theoretical construct, describing the internal processes of information processing that cannot be observed directly*"[52] and therefore needs to be estimated using methods that can generally be divided into subjective and objective measurements as well as according to causal relationship.[52] Typically, the subjective methods are based on questionnaires for retrospective self-reporting after the learning task has been completed whereas the objective methods are applied concurrently.[53] The objective methods can be based on physiological parameters such as pupil dilation or other physiological indices, neuroimaging such as functional magnetic resonance imaging (fMRI), or the dual-task approach.[52] The latter consists of adding a secondary task that needs to be performed simultaneously with the primary learning task. Examples include tapping a rhythm,[54] mental arithmetic tasks,[55] and reaction time.[56]

CL has not been extensively studied in the context of SBT. Secondary task reaction time measurement has been found to be sensitive to changes in CL during SBT of knot-tying;[57] subjective methods for CL measurement had acceptable correlation with performance measures (time and number of movements) in VR SBT of salpingectomy;[58] and haptic feedback during SBT reduced CL especially for experienced surgeons.[55]

In this thesis, the secondary task method is used to estimate CL and measured reaction time in response to a visual or auditory cue. This involved series of repeated measurements both at baseline (to establish individual mean reaction time before and after the simulation) and during the simulation and therefore report relative changes in reaction time as an estimate of changes in CL from a non-active to an active stage. A major limitation of the dual-task approach is that the different components of CL cannot be differentiated.[50]

In their 2005 and 2019 reviews, van Merriënboer, Sweller and Paas highlight several important areas of current CL research: complex and real-life learning tasks, lengthy training programs, and the effect of increasing expertise on which instructional methods that work well (the expertise reversal effect).[49,50] A challenge is that principles that are found useful for simple task experiments yield poor results in complex motor skills tasks.[59] Similarly, short laboratory experiments fail to consider the effects of for example motivation during extended training programs.[49] Finally, as the novice gains experience, there is a need for an instructional design that includes more authentic training tasks, deliberate practice activities, supports for self-regulated learning, tailored feedback, and repeated practice.[60]

This motivates exploring CL in this thesis because SBT of temporal bone surgical skills represents a complex learning task situated in a distributed training program with focus on self-regulated learning.

1.6 Clinical imaging of the temporal bone and segmentation

Conventional CT has for many years been used for clinical imaging of the structures housed within the temporal bone because it is well-suited for visualization of the ossicles, the bony canal of the facial nerve, and the bony cavities of the inner ear including the cochlea.[61] However, the ionizing radiation limits its use for routine imaging. Magnetic resonance imaging (MRI) can on the other hand be used to visualize fluid filled cavities such as the cochlea and vestibular aqueduct, soft tissue structures such as the vestibulocochlear nerve, and soft tissue pathology such as cholesteatoma.[61] Nevertheless, MRI is time consuming, resolution is lower than CT, and bone is not well presented. This altogether precludes the use of MRI for VR simulation of the temporal bone.

Cone-beam CT (CBCT) uses a cone-shaped x-ray beam for digital volume tomography of a smaller anatomical area.[62] This reduces exposure time and radiation dose while providing high resolution images. CBCT has mainly been used in dentomaxillofacial imaging but is well-suited for imaging of the paranasal sinuses and the temporal bone.[63,64] Most temporal bone structures and key surgical landmarks are equally well visualized with CBCT imaging compared with conventional CT.[65,66] Unsurprisingly, there is increasing interest in using CBCT in the routine evaluation of otologic patients and many larger otology centers are implementing

CBCT for "in-office" imaging. Routine clinical imaging including CBCT has several potential applications besides being part of the diagnostic workup: it can be used to guide for example choice of optimal CI electrode pre-operatively,[67] determine electrode placement in post-operative scans,[65] and could also enable patient-specific VR simulation that can be used for pre-operative surgical planning and rehearsal.[68]

VR simulation for patient-specific rehearsal based on clinical imaging needs to provide models that are accurate in the representation of the patient's anatomy and also provide realistic visual cues for the surgeon. The first step in creating such models from clinical imaging is segmentation, i.e., the delineation of relevant anatomical structures and surgical landmarks in the imaging dataset. Consequently, the clinical imaging needs to be of high-quality and with sufficient resolution for the structures to be visible and clearly delineated from surrounding tissue and structures. Altogether, this makes CBCT well-suited for the purpose of patient-specific VR simulation.

Manual segmentation is time consuming and depends on number and complexity of structures that need to be segmented as well as scan resolution because a higher scan resolution results in more slices that need to be manually reviewed.[69] Manual segmentation is in other words not clinically feasible for patient-specific VR simulation for case rehearsal prior to surgery. Guided and semi-automated approaches to segmentation reduces the time for creation of patient-specific VR models considerably but still averages 20–30 minutes per case.[70,71] Fully automated segmentation of temporal bone anatomy is therefore highly desirable and established approaches include statistical shape modelling,[72] atlas-based modelling,[73,74] and neural network algorithms.[75] Recently, we have further augmented the atlas-based approach to also model cochlear microstructures such as the scala tympani and vestibuli.[76]

Once the temporal bone anatomy has been segmented in the clinical imaging dataset, the different anatomical structures and landmarks as well as the temporal bone itself need to be rendered for visualization and interaction in the VR environment. This is no trivial task and involves substantial processing and post-processing in order to provide realistic visual cues in VR simulation.[12] For this to be possible with patient datasets, integration of several processing steps would need to be automated and streamlined in the future.

1.7 Aims and objectives

The observed plateau in learner performance represents a major challenge for VR simulation training of surgical technical skills including in temporal bone surgery. Different strategies to raise or overcome this plateau and optimise SBT can be conceived with the overarching research question: *How do we best deliver simulation-based training for optimal learning in mastoidectomy?* To answer this question, we need to explore different conditions of learning because what we do outside, inside, and in-between simulation matters.

First step is establishing the "*outside*" of SBT: valid and reliable measures of performance that can be used to document the effects of learning interventions in and outside of the simulation environment. Such assessment is the basis for providing direction and feedback during simulation, to motivate learners, and to track progress for advancement within simulation or determine readiness for supervised surgery in the OR.

We also need to study the effects of what we do "*inside*" SBT, i.e., how we present and scaffold the learning task during practice. This requires the application of contemporary medical educational theory and conceptual frameworks to guide and understand the effects of different learning interventions. This for example could include applying the concept of directed self-regulated learning to VR simulation training of mastoidectomy and trying to get a better understanding of the learning processes involved using the framework of CLT.

Finally, we should consider the "*in-between*" conditions of SBT such as how we organize training, how we keep learners motivated and cognitively engaged with learning, and how we ensure progression of learners and relevancy of SBT for learners of all levels. Ultimately, temporal bone

surgical simulation might be transformed from being a tool used mainly for the initial training of complete novices to a platform that is incorporated in to all stages of learning including preparation of the experienced surgeon for any specific patient case in the clinic.

This is no small challenge and requires much development and research. This thesis aims to provide a few of the necessary pieces and expand the current state of VR simulation training and assessment of mastoidectomy by exploring different learning conditions with the following specific objectives:

1. Develop metrics-based automated assessment for VR simulation training of mastoidectomy, define pass/fail performance standards based on the expert performance framework, and investigate consequences of standard setting.
2. Explore reliability of final-product assessment in relation to different training conditions and strategies using generalizability theory (G theory).
3. Investigate the transfer of distributed VR simulation training to cadaveric dissection performance.
4. Compare the CL induced in the VR simulation and cadaveric dissection training to better understand the role of learning environments.
5. Establish the effects on CL of distributed and massed practice conditions.
6. Investigate the effects on final-product and metrics-based performance of intermittent simulator-integrated tutoring.
7. Determine the effects on performance and CL of ultra-high fidelity VR simulation training.
8. Explore clinicians' perceptions of patient-specific VR simulation based on CBCT imaging.

2. Simulation-based assessment

2.1 Background

Assessment of performance is pivotal for establishing the effect of any learning intervention. In temporal bone surgical training, most of the existing performance assessment tools were developed for use in the operating room or to evaluate performance of cadaveric dissection.[34,36,37] With the introduction of VR simulation, we found a need to establish final-product assessment (FPA) in this context and therefore modified the Welling Scale (WS1) final-product analysis tool developed at the Ohio State University.[37] We established that FPA could be used in the VR temporal bone simulation context with a higher inter-rater agreement than in cadaveric dissection most likely due to the standardized virtual temporal bone model.[77] Consequently, we have used this 26-item modified Welling Scale for performance assessment going forward.

Evidence-based pass/fail criteria have not been established for any of the currently developed assessment tools and reliability evidence is also meagre as highlighted in a review on assessment of mastoidectomy performance.[38]

In addition to assessment of performance by human raters, simulation-based metrics can be used for automated feedback and assessment of performance.[39] In VR simulation training of temporal bone surgery, a large number of metrics have been proposed, and in a systematic review, we have mapped the current evidence for these.[78] Substantial validity evidence were found for procedural time and derivatives of time, volume removed per time (efficiency of drilling), force applied near vital structures, and drilling the correct volumes. Other metrics such as angle of drilling, stroke distance and velocity, the burr diameter and type of burr had only some to moderate validity.

Translating these metrics into valid scoring systems and usable feedback is more challenging and in commercial simulators, such assessment is often implemented without any validity evidence,[17] emphasizing the need for more research on metrics for automated, simulation-based assessment.

2.2 Using the expert framework for developing a metrics-based score and standard setting of performance

SBT can rarely lead the trainee to the level of the expert but can be used to increase the competency of the trainee, bringing them from novice to proficient.[20] However, SBT will optimally induce some of the behaviors observed in experts.[79] Such behaviors could in temporal bone surgery for example constitute goal-directed behavior, efficient drilling, and a safe performance.[80] Even though experts perform differently in a VR temporal bone simulation environment than in cadaveric dissection,[81] they do maintain expected behaviors such as drilling more efficiently, exercising caution around critical structures, and achieving proper exposure and overview of the surgical field compared with trainees.[82]

Many of the proposed metrics can be difficult to observe and/or record in real life. Consequently, expert consensus[83] is of limited use in developing metrics-based assessment. Instead, expert performances recorded in the simulation environment can be used for automatic scoring of mastoidectomy performances.[84] Even though this approach has been used for automated assessment in temporal bone surgical training, no study has compared metrics-based scoring with existing assessment tools nor established proficiency levels for established assessment tools or metrics-based scoring systems.

In *paper I*, we recorded VR mastoidectomy simulation performances by experienced surgeons and residents (who were also study participants in *paper III*) for standard setting of assessment and to investigate the consequences of such standard setting. Final products were saved and unsurprisingly, the experienced surgeons significantly outperformed the residents (mean difference = 4.9 points, $p<0.001$). A final-product score (FPS) of 19.5 points was set as a cut-off score for proficiency, resulting in 60 % of the experienced surgeons having a passing performance in their third procedure in contrast to only 5.4 % of the residents.

A total of 129 different metrics and derivatives were recorded but only unique metrics that could discriminate between residents and experienced surgeons (with experienced surgeons having the better performance as well as performance increasing with repetition) were included in a metrics-based score (MBS) model. This resulted in 17 different metrics that were then classified within five components: time and force efficiency, burr size efficiency, burr type efficiency, hesitancy, and goal-directed behavior. Each component was weighted equally, resulting in a combined metrics-based score (ranging from 0 to 100 %). For the MBS, the experienced surgeons outperformed the residents (mean difference = 16.4 %, $p<0.001$). A cut-off score for proficiency of 83.6 % was established, resulting in identical pass/fail rates to that of the FPS. We considered this 60 % pass rate of experienced surgeons reasonable because the experienced surgeons by their third procedure were still on the steep part of their learning curve in the VR simulator. However, data from the experts' learning curve plateau would have been better for standard setting but this was not feasible to obtain.

We further explored the MBS in relationship to other conditions of learning and variables:

A) Repeated and distributed practice increased the MBS following a traditional negatively accelerated learning curve with 50 % of performances passing the established standard after 13 procedures (Figure 3), supporting that the proficiency level is not unobtainable in simulation.

B) The MBS was poorly correlated with the FPS (Pearson's $r = 0.30$) but moderately correlated with the FPS per minute (Pearson's $r = 0.63$). This was unsurprising as the MBS mainly reflects efficiency of drilling whereas the FPS mostly considers volume drilled and violation of structures. Consequently, we found the need to add an additional criterium—sufficient volume removed—because a large percentage of the residents' procedures were insufficiently drilled (mean FPS of 14.0 points), leading to an inflated efficiency (mean of 0.80 points per minute).

Finally, the MBS fails to capture other critical aspects of performance such as not causing injury to critical structures and drilling in the wrong places and should therefore not be used as the only

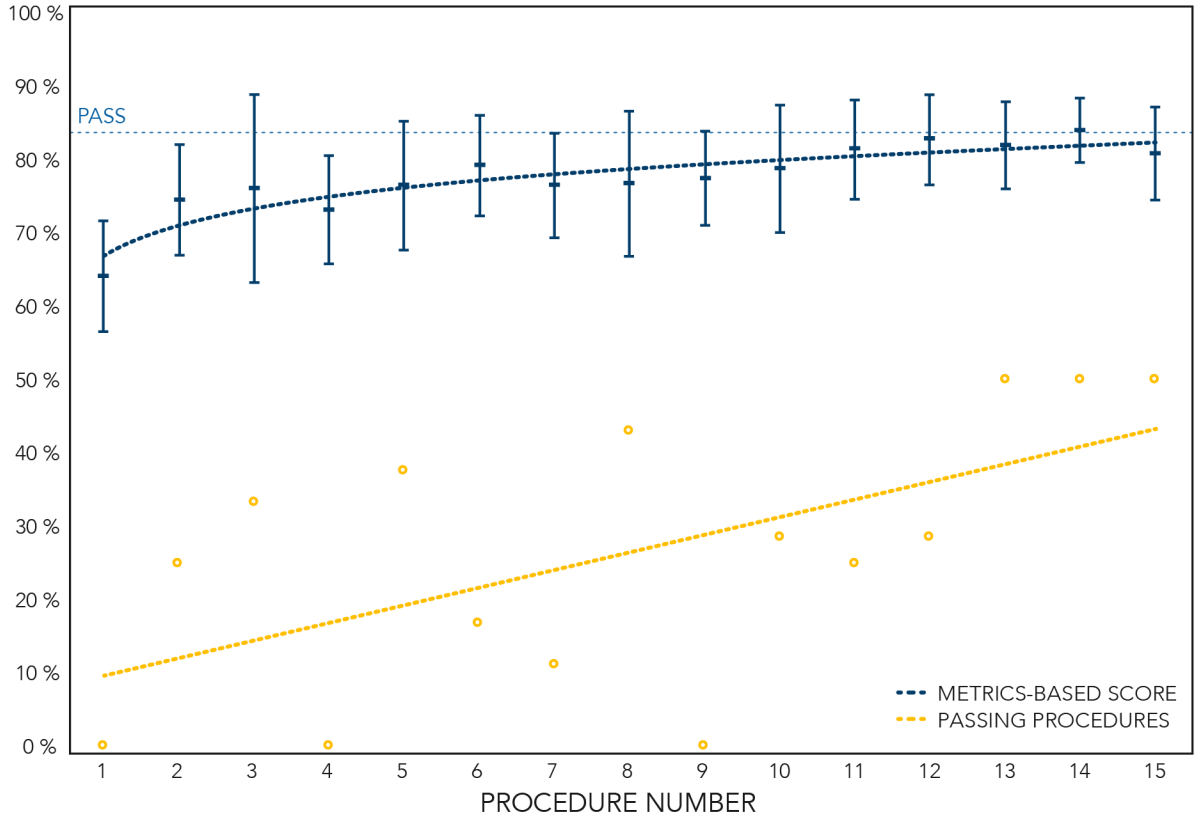

Figure 3. Means plot of the metrics-based score (MBS) and the total number of passing performances. Bars indicate 95 % confidence intervals. The MBS pass/fail standard is set at 83.6 %.

performance outcome. Metrics-based assessment can most likely be improved in relation to this by defining metrics for specific volumes of interest.[85] In contrast, FPA reflects little of the process and integrating both types of assessment seems optimal.

Altogether, this highlights several general issues of simulation-based assessment: First of all, multiple sources of validity evidence need to be considered including relationship to other variables and consequences of standard setting. Excellent discrimination between experienced surgeons and trainees is not sufficient as the only validity argument.[86] Also, several mechanisms are needed to ensure that a safe performance is learned in self-directed VR simulation training. Key elements could be feedback and learning supports that facilitate cognitive processes. In agreement with the expert performance framework, observation of expert performances can be used to inform instructions, feedback, and other learning supports for novices that facilitate learning expert behaviors.

2.3 Exploring the reliability of simulation-based assessment

Assessment is typically limited by the assessment task, assessment conditions, and the number of observations, resulting in the reliability of the assessment requiring extensive consideration.[87] Generalizability theory (G theory) can integrate different sources contributing to performance variability and measurement error and be used to explore reliability.[88] In simulation-based assessment, reliability is often reported as the number of observations needed to achieve a specific Generalizability coefficient—typically a G-coefficient >0.8.[89] In temporal bone surgical training, a G-coefficient of 0.64 for the original Welling Scale could be achieved using two raters observing two cadaveric temporal bone dissection performances by each trainee.[90]

Despite simulation-based assessment representing a relatively controlled and reproducible

environment, contextual factors could potentially affect the Generalizability coefficient.[91] This has received little attention in the literature and we therefore wanted to explore the consequences on reliability of different contextual variables.

In *paper II*, we pooled all the assessment data obtained in the period 2012–2018, representing >3,500 final-product assessments of temporal bone performances under different practice conditions. We found item difficulty to contribute to the largest source of variance, with little variance being introduced by raters and an acceptable ability of the assessment tool to discriminate between a high and a low performing learner. With increased experience of the learner (medical student, resident, and experienced surgeon) more observations were needed to achieve reliable assessment. More observations were also required for VR simulation performances compared with cadaveric dissection performances (Figure 4, left). However, for the VR simulation performances, fewer observations were needed as the graphic fidelity increased (Figure 4, right). Practice organization (massed and distributed practice) and simulator-integrated tutoring had only little effect on the number of observations needed for reliable assessment. In contrast, more observations were needed if the learner was on the initial part of the learning curve compared with the plateau phase. Altogether, the Generalizability analysis adds to the reliability evidence of FPA of mastoidectomy performance and more generally, illustrates how contextual variables can have large effects on reliability. This cautions the use of single-study data for generalizability of assessment in other contexts. In other words, reliability outcomes from one study cannot be assumed to hold in other contexts, which especially for high stakes assessment warrants specific considerations on reliability.

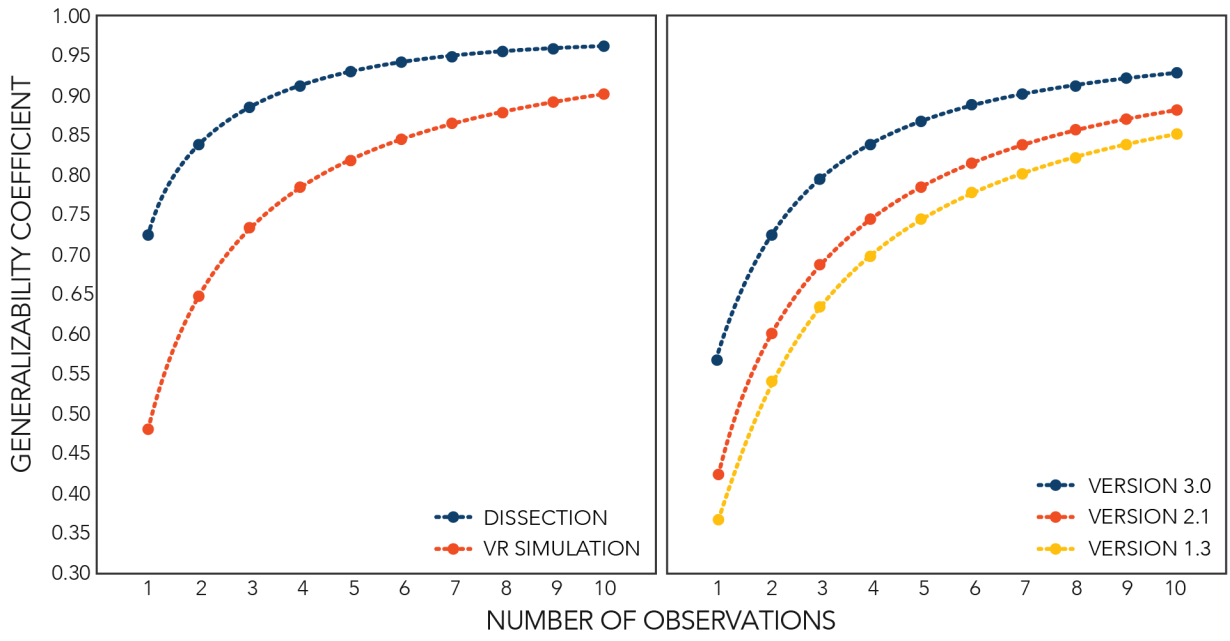

Figure 4. Examples of projections on reliability (D-studies).

3. Training organization

3.1 Background

In the motor skills learning literature, it is well-established that distributed practice is superior to massed practice.[42] This has also been corroborated for surgical skills training for shorter procedures such as microanastomosis and basic VR laparoscopy skills (transfer of a simple object).[92,93] Mastoidectomy represents a more complex surgical procedure and supporting knowledge on shorter and simpler procedures, we have previously demonstrated that distributed VR simulation practice results in a significantly higher end-of-training performance than equal amounts of massed practice (mean difference 2.7 points out of 26 points), p=0.002)[16] as well as a better retention at three months post-training mainly reflected in improvement of time used.[94]

From an organizational point of view, it seems efficient to gather educators, trainees, and expensive equipment. Indeed, intensive learning events with massed practice over one or two days are still common in ORL and temporal bone surgery despite evidence of poorer educational outcomes. Boot camp concepts with simulation-based skills training of a variety of ORL procedures over the course of a single day have gained considerable popularity in North America.[95,96] Similarly, in our survey of European ORL training departments, we found that 75 % of responding departments send their trainees on 1–3 day long temporal bone courses.[3]

An alternative to these intense, massed practice courses is to offer learners distributed VR simulation training prior to the wet temporal bone course. However, the effect on cadaveric dissection performance (i.e., transfer) and CL in relation to learning environments and practice organization has not previously been investigated for mastoidectomy.

3.2 The effects of distributed VR simulation training on transfer

We have previously established that three hours of VR SBT before cadaveric dissection training of mastoidectomy can improve dissection performance by 52 %.[97] We also know from the learning curves that the optimal training program would consist of distributed practice blocks each consisting of three repeated procedures.[16]

In *paper III*, we wanted to investigate the effect on cadaveric dissection performance (transfer) of the suggested structured, distributed VR simulation training program. Participants were residents attending the national temporal bone course, who we invited for additional VR simulation training of mastoidectomy in the three months leading up to the temporal bone course. Nine out of 37 trainees from the 2016 and 2017 national temporal bone courses accepted the additional VR simulation training and each trainee completed five training blocks of three identical procedures before the course (intervention). The remaining course participants served as controls and both the intervention and control participants received the standard three hours of VR simulation training during the temporal bone course.

As expected, end-of training VR simulation performance of the intervention group (who completed a total of 18 procedures) was higher than the control group (who completed three procedures during the temporal bone course) with a mean difference of 4.4 points (95 % CI [1.8–7.0], p<0.005) (Figure 5, left side). More importantly, the effect of distributed VR simulation practice transferred to the cadaveric dissection setting, resulting in the intervention group outperforming the control group with a mean difference of 2.6 points (95 % CI [0.7–4.4], p<0.005) (Figure 5, right side). In other words, the distributed VR simulation training program

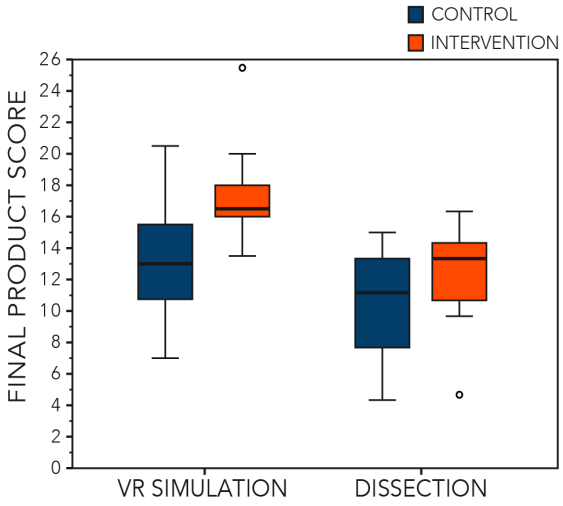

Figure 5. Box plots of final-product performance.

further increased cadaveric dissection performance 25 % compared with just a single block of VR simulation training.

In contrast to our previous study where time for each repeated procedure was fixed at 30 minutes,[16] we allowed participants unlimited time for each procedure. This resulted in the final-product performance per minute to best reflect the combined increase in performance because the time needed to complete the procedure was substantially reduced with repeated practice.

Overall, we demonstrated that distributed practice has an additional positive effect on subsequent cadaveric dissection performance compared with the standard amount of VR simulation practice during our temporal bone course. We found the rate of participation to be high (24 %) considering the geographical distribution of the trainees and that trainees were not compensated for the time they used for the additional VR simulation training at the simulation centre. Further, with local and at-home training being a feasible option,[98] implementing distributed VR simulation training before the temporal bone course can augment the otherwise very limited opportunity for training on human temporal bones during the national temporal bone course.

3.3 The effects of VR simulation training conditions on CL

CL in SBT has received increasing interest over the last two decades.[53] However, at the time of our first studies, CL was only scarcely studied in SBT of surgical skills and procedures in contrast to full-scale simulation of medical scenarios. Since our first studies, other studies have added that reducing task complexity in SBT of lumbar puncture decreased CL but that this had a negative effect on transfer of performance,[99] and that CL can be used as predictor in differentiating levels of experience in SBT of ultrasound skills.[100] One of the main gaps, we identified in the literature, were whether practice conditions affect CL.

In *paper IV*, we therefore investigated CL in VR simulation and cadaveric dissection training of mastoidectomy (in extension of the previously mentioned study[97]) and compared the CL imposed on trainees under these two conditions. We used a cross-over design to balance out effects of repetition and did serial measurement of secondary task reaction time. One cohort of 20 temporal bone course participants (residents) received VR simulation training before cadaveric dissection and another cohort of 20 course participants cadaveric dissection before VR simulation training. We found that the mean relative increase in CL during VR simulation training was 1.20 (95% CI [1.18-1.22] compared with 1.55 (95% CI [1.52- 1.59]) in cadaveric dissection ($p<0.001$). In other words, cadaveric dissection induced substantially more CL on the trainee than the VR simulation (task and environment combined).

A limitation of the study was that we measured reaction time using two different systems: in VR simulation, an automated system presented a visual letter cue on-screen and participants had to respond by pressing on the corresponding key on the keyboard; in cadaveric dissection, a manual system with an auditory beep cue required the user to respond by pressing a foot switch. This could potentially result in different speed-accuracy trade-offs even though we could not verify any differences in precision in the automated system with the visual cues.[101]

In *paper V*, we wanted to investigate the effects on CL of distributed VR simulation training. As part of another study,[16] one cohort of 21 medical students completed distributed practice and another cohort of 19 medical students that completed massed practice. Both cohorts performed 12 identical mastoidectomy procedures in the simulator. In the distributed training program, participants were allowed to perform two procedures per session with at least three days between sessions. The massed training program consisted of performing all the procedures successively in one day. Each cohort was further randomized for simulator-integrated tutoring during the first five procedures or no simulator-integrated tutoring at all. CL was estimated using the automated visual-cue secondary task system in the VR simulator.

We found that the distributed practice program linearly and significantly reduced CL about 15 %. In

contrast, CL remained largely unchanged under massed practice conditions. Simulator-tutoring did not seem to affect CL and the effect could therefore be attributed to the distribution of practice blocks. This finding is in agreement with the motor skills literature, which finds that consolidation of memory is time-dependent,[102] substantiating how the positive effects of distributed practice result from actual learning.

4. Simulator-integrated tutoring

4.1 Background

Feedback is an integral part of motor skills learning.[42] In surgical skills training, Reznick describes feedback as part of "*cognitive apprenticeship*" and lists steps that support this as modelling, coaching, scaffolding, articulation, reflection, and exploration.[1] Unsurprisingly, feedback is an effective feature of SBT[44] but needs to be considered as it might also negatively impact learning of procedural skills.[103] The principles of DRSL also emphases providing direction and guidance.[26,27] Altogether, there is a need to study feedback in VR simulation training of mastoidectomy and in the virtual learning environment, feedback might be automated but the appropriate way to deliver this to the learner needs research.

Unique for temporal bone surgery in contrast to most other surgery, is the precision with which we in simulation can record, save and measure the removal of bone during a mastoidectomy procedure. This allows defining the volume that needs to be removed during the procedure as a reference for feedback and assessment of performance. Defining this volume is an example of how capturing the expert performance[79] can be used to establish the learning parameters for SBT.

The reference volume can for example be presented to the learner as summative feedback/assessment at the end of procedure (i.e., how much of the drilling was inside/outside of the reference volume).[17,77] In our systematic review of performance metrics for mastoidectomy, reference volume metrics were found to have substantial validity evidence.[78]

Another use of the reference volume is to concurrently green-light the volume to be drilled during the procedure.[104] This is an intuitive and visual way of instructing the learner compared with traditional approaches such as text, illustrations, and videos, where the learner has to extract the information and then apply them during the drilling. The simulator-integrated tutor-function constitutes a learning support for concurrent feedback in the directed, self-regulated learning environment.

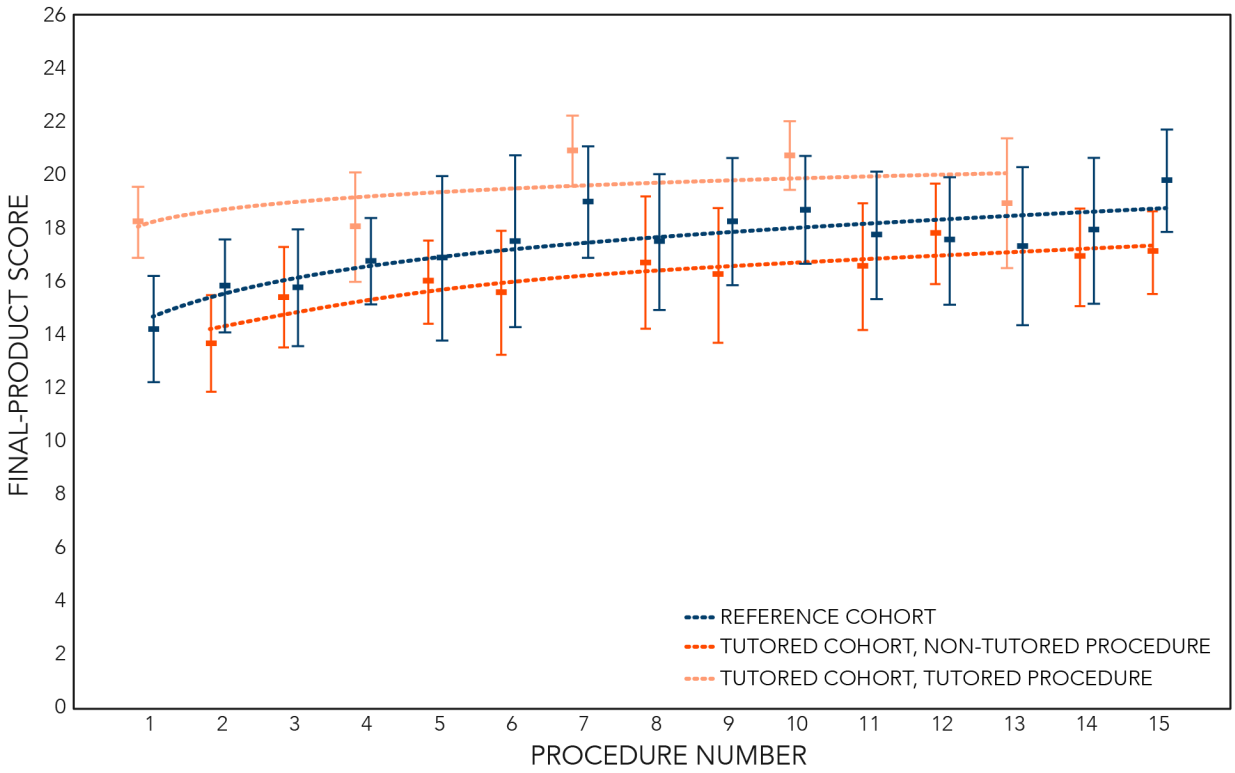

Figure 6. Learning curves of tutoring. Means plots with 95 % confidence intervals of final-product performance for the tutored cohort, according to tutored and non-tutored procedures, and the never-tutored reference cohort.

In our previous work, we have investigated the role of the simulator-integrated tutoring in relation to the learning curves of repeated performance.[16] Tutoring was unsurprisingly found to increase performance when active. However, performance dropped markedly when the tutor-function was discontinued, and several untutored repetitions were needed for performance to catch up with the performance of never-tutored learners. Interestingly, we found that distributed practice offered some protections against this tutoring over-reliance effect.[16]

4.2 The effects of tutoring on performance

In *paper VI*, we therefore wanted to explore the effect of simulator-integrated tutoring in more details and to investigate if using tutoring only intermittently would alleviate the negative effects of tutoring and increase the end-of-training performance. The study population consisted of medical students to have complete novices: one cohort of 16 participants practiced with stepwise instruction texts and images and in every third procedure additional color-coded simulator-integrated tutoring in a distributed training program. Another cohort of 14 participants served as a reference and practiced using identical learning conditions including the stepwise instructions but never received tutoring. This enabled us to study both the immediate effects of tutoring on performance (i.e., effects while the tutor-function was turned on) as well as effects of intermittent tutoring on the cohort's non-tutored procedures (i.e., effects on learning). Outcomes were final-product performance as well as simulator metrics.

For the final-product performance (Figure 6), tutoring had a large and positive effect for the procedures where it was turned on (mean difference 3.8 points out of 26 points, $p<0.001$) similar to previously found. In contrast to expectations, this did not translate to a better performance in subsequent non-tutored performances: overall, the reference cohort consistently performed superiorly to the intervention cohort (mean difference 1.4

points, $p<0.001$). However, further analysis substantiated that this negative effect on learning was confounded by not removing enough volume: some items could not be positively rated because too little bone over the critical structures was removed to assess if the performance was "safe" and rate it positively.

For the performance in relation to simulator-metrics, tutoring resulted in a higher metrics-based score (MBS) and more procedures passing the standard set for the MBS, reflecting increased efficiency of drilling compared with the reference cohort. For individual metrics, the effects of tutoring on subsequent non-tutored performances were mixed: tutoring led to fewer collisions with the incus and the inner ear structures but also to more collisions with the ear canal skin and tympanic membrane as well as drilling more bone not directly visible (obscured).

Overall, we found that intermittent tutoring in a distributed practice program failed to increase performance beyond the learning curve plateau, and even though there were a few positive effects on learning, this tutoring over-reliance effect was not mitigated. Continuous feedback by this sort of tutoring has only few positive effects that are largely outweighed by the negative effects on learning of the mastoidectomy procedure. Our findings are, however, limited in relation to participants, data collection, and outcome measurement, which is further unfolded in chapters 7.2 and 7.3. Altogether, tutoring in VR simulation training of temporal bone surgical simulation should be used with caution and might best be suited to introduce only the very first procedures to complete novices.

5. Fidelity of simulation

5.1 Background

Fidelity of the simulation in relation to learning is much debated—how realistic does simulation need to be for learning? A recent systematic review found that for procedural skills training, "*skill after training with low fidelity simulators was not inferior to skill after training with high fidelity simulators*".[105] This conclusion might seem somewhat "inconclusive" in determining if fidelity affects learning but also highlights the problem with simulator fidelity typically being classified as "high" or "low". This dichotomization is problematic because fidelity can be considered for example in relation to visual, tactile, or functional features of the simulator and especially for educational simulation, the effect of fidelity is complicated by learning context and interactions.[106] Hamstra and colleagues therefore suggest abandoning the term fidelity and instead consider physical resemblance and functional task alignment, with focus on the latter in establishing educational effects especially in order to enhance transfer of learning.[106] In other words, the agenda should be to design simulation as a mean to improve learners' real-life abilities and to hone in on the elements—whether physical or otherwise—that actually contribute to this.

For surgical skills training, physical resemblance of the simulation might be important as haptic and visual cues are required for a safe surgical performance. Systematic reviews have established haptic feedback in surgical simulation to improve the training effect[107] and that 3D stereovision in laparoscopy seems to improve speed and reduce errors compared with 2D vision.[108]

Less is known on the topic of physical resemblance in relation to learning in temporal bone surgery, but all major VR temporal bone simulators integrate haptic simulation as the drilling experience is a key element of the surgery. Further, in real-life temporal bone surgery, a surgical microscope is used to enable the surgeon to visualize the minute anatomy of the temporal bone and especially the fine visual cues such as the vascularization of underlying structures through a thin layer of bone. In *paper II*,

our data suggest that the latter is important for the performance during VR simulation of mastoidectomy: as the graphical realism increased for improved visual cues from simulator version 1.3 to 3.0, the adjusted performance increased from 15.5 points (estimated marginal mean, 95 % CI [15.1-15.9]) to 18.3 points (estimated marginal mean, 95 % CI [17.4-19.2]). Nevertheless, more research is needed on the role of simulator physical resemblance and functional task alignment to improve SBT of mastoidectomy.

5.2 The effect of ultra-high-fidelity VR simulation on performance and CL

In *paper VII*, we wanted to explore VR simulation using an eyepiece from a digital operating microscope compared with the standard screen-based projection used in our other studies. This eyepiece represents an increase in both simulator physical resemblance and functional task alignment since it mimics the operating microscope used during real-life surgery while increasing depth perception and clarity of projection. We chose to use the term "ultra-high-fidelity VR" (UHF VR) for this condition to separate it from conventional screen-based VR (cVR), which is also considered high-fidelity.

We recruited 24 medical students and used a cross-over design to determine the effect of the two conditions (UHF VR vs. cVR) and their order. Participants had previously received eight hours of cVR simulation training related to CI surgery and were therefore well-acquainted with the Visible Ear Simulator. They were randomized to complete two partial mastoidectomies (up until the point of posterior tympanotomy) in cVR followed by two identical procedures in UHF VR, or vice versa, and allowed 45 minutes for each procedure.

To investigate the effect on performance (primary outcome), we assessed final-product performances using a modified Welling Scale reduced to only the 17 items relevant to the partial mastoidectomy. To explore the effect on CL, we measured secondary reaction time manually using a reaction timer (i.e., as the response time on a foot pedal to a beep). Measurements were done in repeated series and

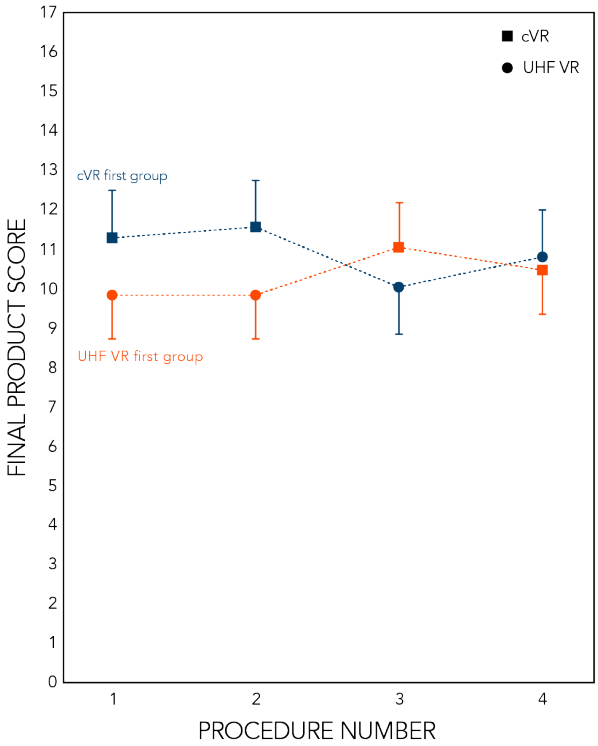

Figure 7. Mean plots of final-product performance.

calculated relative to baseline measurement.

We found that performances in UHF VR were significantly lower than those in cVR (mean difference 1.0 points, p=0.02) and that this was not affected by order (Figure 7). However, participants who were unable to achieve stereovision in UHF (despite our best efforts to optimize their view in the oculars) performed significantly worse regardless of condition.

The group that received UHF VR simulation training first demonstrated a significantly higher CL increase relative to baseline than the group that received cVR first (28 % and 18 %, respectively). Adjusting for the order, the UFH VR condition still induced a 5 % higher CL than cVR (p=0.03). The ability to achieve stereovision in UHF VR only trended to impact CL.

Altogether, this suggest that even for trained novices familiar with the simulation environment, increased physical resemblance and functional task alignment through the use of an operating microscope eyepiece did not benefit learning as it negatively affected performance and induced more

CL. This corroborates our finding in *paper IV*, where the high complexity of the dissection learning condition including the use of the operating microscope induced a much higher CL than cVR. Ultimately, improved instructional design and feedback might be of more benefit than further increasing realism of simulation, at least for novices and intermediary learners.

6. Patient-specific simulation

6.1 Background

A holy grail in VR surgical simulation is patient-specific simulation, which would enable the surgeon to rehearse and plan surgery for the individual patient in the virtual environment ahead of actual surgery. Potential benefits include increased surgeon preparation and confidence, increased efficiency and reduced operating time, and a reduction in surgical errors and complications.[68,109] In temporal bone surgery, patient-specific simulation could help visualize and prepare for difficult anatomical variants and malformations, tailor treatment and surgical intervention including CI, and plan minimally invasive surgery.[68,110]

Even though patient-specific simulation could benefit trainees as well as experienced surgeons, few options are available in temporal bone surgery and include tablet-based planning tools,[67] 3D-printed physical models,[111] and VR simulation. A review from 2015 on VR simulation for pre-operative preparation in otologic surgery[68] identified only a single study.[71] Since then, another study has also evaluated a VR simulation system for pre-surgical practice and trainees perceive the system to be useful and to increase their confidence.[112] Further, in a comparison with intra-operative recordings, patient-specific simulation was accurate in the evaluation of the round window.[113]

A major limitation of current patient-specific VR simulation of the temporal bone is related to the low graphical fidelity.[71] This highlights the challenge in transforming clinical imaging datasets into usable models for patient-specific VR simulation. The required segmentation of the surgical anatomy in imaging datasets is time consuming and none of the previous systems have used a validated segmentation routine, so accuracy of segmentation is undetermined, which limits clinical use.

The Visible Ear Simulator does currently not feature integrated processing of patient imaging datasets for VR simulation even though new models based on CBCT have now been integrated[12] as a step towards future direct import. In contrast, the OSU

virtual temporal bone system (OSU-TB) was based on digital volume rendering of CT datasets (conventional CT/micro-CT) from its inception, which has enabled easy import of new datasets including clinical CBCT scans. These need to be processed for segmentation of relevant anatomical structures but this has largely been automated using an atlas-based algorithm, for which the accuracy has been validated against manual segmentation.[73,74]

6.2 The potential of patient-specific simulation

In *paper VIII*, we wanted to evaluate the OSU-TB system for patient-specific simulation including the visualization of the automated segmentation and perceived usefulness by trainees and experienced surgeons.

We used 22 pre-operative and de-identified clinical CBCT datasets from adult CI candidate patients with normal anatomy. The datasets were processed using the automated processing pipeline, resulting in segmentation of the facial nerve, chorda tympani, and lateral semi-circular canal, which in addition to the bony structures such as the incus, constitute key surgical landmarks in CI surgery. Nine surgeons of different experience (four attending neurotologists, three fellows, two residents) were recruited from three institutions and performed temporal bone drilling relevant to CI surgery (i.e., mastoidectomy, facial recess and round window approach) on as many of the patient-specific cases in the OSU-TB simulator as their time permitted.

After completion of each patient-specific case in the VR simulator, the participant completed a structured questionnaire on the utility of the simulation including simulator ease of use and reality of different aspects of the simulation, usefulness for presurgical planning, impact on appreciating anatomical subtleties, and perceived impact on trainee learning.

Each surgeon completed an average of 16.5 of the 22 cases and each case was evaluated by at least six different surgeons. The surgeons generally reported favourably on the overall experience and 71.6 % of the simulations were rated highly overall. Most simulations were reported positively in relation to the surgeon better appreciating anatomical subtleties with 10.7 % of simulations being reported as contributing to a significantly greater understanding of patient's anatomy compared to that of standard imaging. These subtleties were especially related to the aeration of the mastoid and the width of the facial recess.

The potential impact of patient-specific simulation for training was also rated positively with 44.5 % of simulations being perceived of benefit at the resident level of experience and 37.7 % at the fellow level, indicating the relevancy of patient-specific simulation even for more experienced surgeons.

There was, however, room for improvement in relation to the simulator and the visualization in some of the patient-specific cases. For example, the feel of the drill was generally rated less favourably and surgeons found that the stapes and round window membrane were often poorly represented. Cases with little cortical bone or without bone over the facial nerve or chorda due to the limited field-of-view in the scan were rated poorly. Another 14 clinical scans were originally obtained but not included in the study due to poor scan quality and blurring/motion artefacts. This emphasizes the importance of adequate field-of-view and scan quality control at time of acquisition.

The main limitation of the study was the few participating surgeons, who were recruited from our own institutions and who might be biased towards favouring simulation. Further, datasets were obtained retrospectively and could not be compared against real-life findings or observations for example from intraoperative video recordings. Finally, the evaluation consisted of surgeons' opinions and perceptions after drilling the cases in the simulator. Nonetheless, our findings provide direction for future improvement of the system and also qualifies the potential of patient-specific VR simulation even in cases of normal anatomical variation. Comparison of patient-specific virtual drilling with post-operative imaging might contribute further validation of the accuracy before clinical evaluation.

7. Discussion

This thesis explores SBT and assessment of mastoidectomy with a focus on VR simulation under the conditions of DSRL where the trainee has to rely on the deliberate design of the learning experience such as learning supports and structure of the training program. This further need to take into consideration individual learner needs, goals and previous experience.

The overarching objective is to ensure that the learning potential of VR simulation is exhausted before the learner proceeds to clinical practice. A major challenge in temporal bone surgical training is the apparent performance plateau in the simulation environment, where there is still room for further learning and optimization of training.

This thesis presents research related to the outside, inside and in-between conditions of VR simulation practice in temporal bone surgery to expand our current knowledge and set future direction for VR simulation as not only learning tool for novices but also as a valuable tool for experienced surgeons. In the following, I will attempt a birds-eye view on the main findings and implications, a comparison with current literature and a discussion of general limitations. This will leave many specific details and discussion points to the discussion found in the individual studies.

7.1 Summary of main findings and implications

First, we wanted to further the field in relation to valid and reliable measures of performance in mastoidectomy. This "*outside*" of simulation is important as accurate assessment is the basis for feedback and certification of skills both in SBT and in the clinical setting. Current assessment tools have limited validity evidence[38] and/or require time-consuming observation and rating of performance.[34–37] We therefore wanted to explore performance assessment based on simulator metrics (*paper I*), which can be used for automated and integrated feedback and assessment. Using the expert performance framework,[79] we established pass/fail standards for proficiency-based training for both the new metrics-based assessment (MBS) and the well-established final-product assessment (FPS) and also explored consequences of standard setting. However, our metrics-based assessment failed to capture important aspects of a safe performance whereas final-product assessment reflects a safe performance but little on process and efficiency. There is therefore a need to implement instructions, feedback, and assessment features that supports learning how to perform the procedure safely.

Next, we pooled almost a decade's worth of mastoidectomy assessments (*paper II*) and explored projections in reliability using G theory.[88] This adds much needed validity evidence for our modification[77] to the original final-product assessment tool[37] but also led to some other insights: context variables and learning conditions affect reliability more than is mostly considered in the literature.[114] We found that learner level, simulation fidelity, and learning curve affected reliability markedly, whereas organization of training and tutoring only had marginal effects on projections in reliability. Consequently, reported reliability for a surgical skills performance assessment tool should not be assumed a general trait because it seems closely linked to assessment context and learning conditions. This cautions generalization of reliability coefficients reported in the literature and warrants dedicated reliability studies whenever an established assessment tool is used in a different context. This is especially critical in high-stakes assessment such as for certification but also other consequential decisions such as when the trainee is ready to progress from SBT to clinical practice.

Traditionally, basic temporal bone surgical skills have been acquired on human cadaveric temporal bones, which are an increasingly scarce and costly resource.[3] They should therefore be reserved for refinement of skills after other training modalities such as VR simulation have been exhausted. From an educational perspective, dissection training introduces several important layers of complexity necessary to bridge SBT to clinical training but this complex learning condition might also place high cognitive demands on the trainee.

An important research agenda is therefore to optimize training leading up to dissection training

and ensure a minimal loss of skills due to transfer processes. Distribution of practice over time is well-established in the motor skills literature as superior to condensated and massed practice.[42] This "*in-between*" of SBT is central because motor skills consolidation is time dependent.[102] We therefore investigated the effect of distributed practice on transfer to the cadaveric dissection environment (*paper III*). We found a significant increase in dissection performance after structured and distributed practice compared with the standard three hours of VR simulation practice immediately before dissection at a traditional temporal bone course. This adds to current knowledge that even complex psychomotor skills, such as those required for mastoidectomy, transfer to a considerably increased performance in a higher complexity learning environment (i.e., dissection).

That the dissection learning environment is challenging for the novice learner was further explored with the lens of CLT: according to CLT, cognitive demands that exceeds the capacity of the learner are unfavourable for learning.[48] We therefore set out to investigate CL in the learning conditions of VR simulation and cadaveric dissection (*paper IV*). We found that dissection training induced a significantly higher CL in novices than the VR simulation training. Consequently, we wanted to explore whether distributed VR simulation practice could be used to reduce CL (*paper V*). For novices, we found that CL was significantly reduced at the end of distributed practice whereas it remained unchanged at the end of massed practice. This suggests that VR simulation training not only has positive effects on motor skills performance but also on cognitive aspects of learning most likely through the formation of more efficient mental schemata that are formed over time.[51] The implication of this is that complexity of learning conditions should be considered in surgical skills training.

The early learning curve plateau observed in VR temporal bone simulation training[16,17] is a problem because distributed VR simulation training is only feasible if it can be self-directed, enabling practice at individual trainee's convenience without requiring the presence of human instructors for feedback and guidance. The principles of DSRL emphasizes providing the learner with learning supports for direction, guidance and feedback to successfully scaffold the learning process.[26,27] This requires deliberate design of the simulation learning experience with a consideration of the "*inside*" of simulation to determine the optimal conditions for learning.

The Visible Ear Simulator has an intuitive and integrated tutor-function that can be considered a source of concurrent feedback because it visually and step-by-step guides the trainee to drill the reference volume of the mastoidectomy procedure. We have previously found that the continued use of simulator-integrated tutoring was problematic for learning but also that distributed practice seemed to mitigate some of the negative effects.[16] Therefore, we hypothesized that intermittent tutoring could improve skills acquisition without the negative effects of tutoring over-reliance (*paper VI*). Nevertheless, intermittent tutoring resulted in poorer performances in non-tutored procedures compared with performances of never-tutored participants. Further, tutoring did not consistently improve safety aspects of the procedure such as collisions. This indicates that tutoring degrades motor skills learning and consequently, concurrent feedback such as tutoring during SBT of surgical technical skills should be used with caution.

Another condition of SBT is the physical resemblance and functional task alignment (i.e., aspects of fidelity).[106] Current VR simulators project the simulation onto a computer screen (with/without 3D stereovision) in contrast to cadaveric dissection and real-life surgery, where an operating microscope is necessary for magnification of the surgical field. We therefore introduced the eyepiece from a digital operating microscope into the VR simulation for an increased physical resemblance and functional task alignment (*paper VII*). We found that this did not benefit performance or CL of advanced beginners: this "ultra-high fidelity" VR simulation resulted in a significantly lower final-product performance compared with conventional screen-based VR simulation and also induced a significantly higher CL in the learners. In other words, less complexity of the temporal bone surgical

simulation seems better for novices and how to best bridge simulation to real-life conditions requires further research.

All of the studies in this thesis so far have used the same virtual temporal bone model for training. However, one of the features of SBT associated with positive effects on learning is case variation that can present the learner with a range of difficulty.[44] This is important to support the progression from novice to proficient[20] and simulation might be more valuable to experienced surgeons if it can be used for patient-specific simulation. Patient-specific simulation allows planning and rehearsal ahead of surgery tailored to the individual patient based on clinical imaging and this holds great potential in temporal bone surgery.[68] However, patient-specific simulation is reliant on accurate segmentation of key surgical anatomy in the clinical imaging dataset, which is no trivial task.

Recent developments of automated image processing using an atlas-based approach for temporal bone segmentation[73] allowed us to explore the potential of patient-specific simulation in the OSU temporal bone simulator (*paper VIII*). We found that participating surgeons had a positive attitude towards the patient-specific VR simulation and found that it supported their understanding of the individual patient's surgical anatomy. Further, surgeons perceived patient-specific simulation valuable in the training of residents and fellows. The patient-specific simulation is highly dependent on high quality clinical imaging and limitations were mostly related to narrow field-of-view and artefacts due to patient motion during scan acquisition. To unfold the clinical potential of patient-specific simulation for training and surgical planning, further validation is needed as well as methods for automated segmentation in cases of abnormal temporal bone anatomy such as vestibulocochlear malformations.

7.2 Comparison with current literature

An advantage of SBT and especially VR simulation is the possibility of automated assessment based on objective metrics with much more ease than tracking for example instrument movement during actual surgical procedures.[115] Many commercially available VR simulators offer automated feedback and assessment based on an analysis of performance metrics such as time, path length etc. The selection of these metrics for assessment and feedback is, however, often not based on evidence at time of programming and are in best cases determined by manufacturers in collaboration with medical educationalists and clinical experts and later supported by research into validity.

The field of VR temporal bone surgical simulation differs slightly from this because the available simulators were developed in academic environments and only one simulator has since been commercialized. Performance metrics have therefore been considered from the very beginning to track applied force on the drill near critical structures and to ensure that a safe performance is learned.[6,116] A range of metrics relevant to mastoidectomy have been described in the literature and are supported by varying degrees of validity evidence.[78] Nonetheless, implementation into the VR temporal bone simulators is limited and supporting validity evidence according to a contemporary validity framework was prior to *paper I* largely absent from the literature.[38]

In contrast, there is reasonable validity evidence for (manual) structured assessment of mastoidectomy performance using different assessment tools in the context of intra-operative, cadaveric dissection, or VR simulation training.[38] Nonetheless, for none of these assessment tools, evidence-based pass/fail levels or cut-off scores for proficiency had been determined. This is necessary for the implementation of mastery learning.[22]

Performance standards can be defined using a number of accepted methods such as Angoff (item-centered), Hofstee (whole-test-centered), and contrasting groups (examinee-centered).[117] In a recent systematic review, we found that the most commonly used methods for standard setting in SBT of surgical skills were based on the measurement of performance of groups of learners (i.e., examinee-centered) including experienced learners/experts.[118] Correctly defining the group of experienced learners is therefore of great importance but actual experience varies considerably between studies, ranging from 30 to 4,000 procedures.[118] Another

challenge is that the "*experienced*" group might have the required real-life surgical experience but this might not transfer immediately to the simulated setting. This was exemplified in *paper I*, where the experienced surgeons demonstrated a steep learning curve in VR simulation. This impacts standard setting and might set the pass/fail standard too low.[119]

Reliability evidence for temporal bone surgical performance assessment consists mostly of studies on interrater reliability.[38] In contrast to classical test theory, G theory can take different contributors to variance and measurement error into account.[88] In medical education these facets might be learner, case difficulty, assessor leniency, etc. This enables a more detailed analysis of reliability but G theory is underemployed in surgical technical skills assessment and especially in simulation-based assessment as highlighted by our systematic review.[114]

For FPA of temporal bone surgical performance on cadaver and plaster models, Fernandez et al. have previously explored reliability using G theory:[90] a majority of the variance in performance was attributed to the inconsistent performance of learners. However, the number of performances included in the study was small and the raters were potentially confounded in the study design. In *paper II*, we pooled final-product assessment from a number of previous studies to include as many observations as possible to strengthen our G analysis and study the impact of different learning conditions. We found that most variance could be attributed to item difficulty and interaction between learners and items (item specificity). Item was not included as a facet in the G analysis by Fernandez et al. and was likely encompassed in the person variance. Ultimately, they conclude that using two raters assessing two performances per participant would yield satisfactory reliability of cadaveric dissection performance assessment whereas we find two observations (for example two raters assessing one performance) sufficient. This is on par with what is generally found in the literature.[114]

From different motor skills domains, it is well-established that repeated practice is necessary for psychomotor skills acquisition. For SBT in medical education, a dose-response relationship between amount of training and standardized learning outcomes has also been demonstrated.[120] Unsurprisingly, repeated practice is more efficient if spaced over time—so-called distributed practice—and we have previously demonstrated that this is also true for a highly complex surgical procedure, i.e., mastoidectomy.[16] Nevertheless, much temporal bone surgical training including our national temporal bone course and many similar courses throughout Europe[3] are organized as single-instance, intensive courses with massed practice. This might be perfectly fine if the learning goal is just an introduction to the mastoidectomy procedure for basic knowledge. In that case, distributed VR simulation can be used to enhance skills prior to dissection to make best use of the cadaveric specimens. On the other hand, if the goal is to train surgeons to proficiency in the procedure, extended and distributed practice to an evidence-based level in simulation before supervised surgery should be implemented.

It is therefore important to establish transfer from SBT and a meta-analysis by Lui and Hoy has verified the positive effect of VR simulation training on mastoidectomy performance in cadaveric dissection.[121] However, *paper III* was the first to investigate the added effect of distributed VR simulation training on transfer, adding an important piece of information: namely that not only does performance transfer from VR simulation but increased amounts of practice has additional benefit. Ultimately, that VR simulation training transfers to OR performance remains circumstantial. Even though a recent study has explored distributed VR simulation training before supervised surgery in mastoidectomy, poor methodology—especially the lack of a control group—makes inference on transfer impossible.[122]

CL is little studied in simulation across domains and even less in simulation for health care professional education.[53] Measurement of CL is not straight forward because CL, as described in the introduction, is a construct that needs to be indirectly measured with subjective or objective

methods.[52] In their review, Naismith and Cavalcanti further find that "*inconsistent correlations between CL and learning may be related to issues of validity in CL measures*" and recommend triangulation of CL through by different measurements.[53] Altogether, there is limited and scattered evidence related to the role of CL in SBT and many different methods have been used across studies.[53]

Different learning conditions might modify CL: they can increase CL due to information overload or decrease CL due to better structuring of instructions for easier cognitive processing.[103] In *paper IV* and *paper V*, we presented data from measurement of secondary task reaction time to estimate the relative increase in reaction time compared with individual baseline measurement as a proxy for change in CL. We find that the increased complexity of the learning conditions (task and setting) during cadaveric dissection induces a CL in novices that is considerably higher than during VR simulation, and that repeated VR simulation reduces CL. We later added that distributed VR simulation practice can be used to reduce CL in subsequent cadaveric dissection.[123] The increased complexity of using the operating microscope eyepiece in VR simulation in *paper VII* also increased CL compared with standard conditions.

Nonetheless, establishing causality between CL and performance is difficult and even well-designed studies can only speculate that CL might be a mediator.[99] Further, a high CL can be beneficial if it is within the cognitive capacity of the learner and constitutes germane load.[51] Determining the specific components of CL being modified by learning interventions is also difficult but would really enhance our understanding of the underlying cognitive mechanisms. However, the components are impossible to separate through the secondary task methodology we used and require other methods. Ultimately, the construct and measurement of germane load remains debated.[124,125]

Feedback during SBT is one of the features that is associated with positive effects on learning outcomes.[44] This has also been found in VR simulation training of mastoidectomy where automated feedback messages related to technique and proximity to vital structures[126] or tutoring through volume greenlighting[16] increases performance. Unfortunately, this does not seem to translate into actual learning because performance was found to drop markedly after discontinuation of the feedback even though distribution of practice mitigated some of the negative effect.[16] As demonstrated in *paper VI,* using the tutoring intermittently did not result in better learning. Altogether this corroborates the problem with concurrent (on-going) feedback that was also found in a review on feedback in simulation-based procedural skills training: *tutoring over-reliance* can result from concurrent feedback and terminal feedback seems to produce better long term learning.[103] Tutoring over-reliance can be understood through the guidance hypothesis of the motor skills learning framework: learning is degraded by continuous feedback during the skills acquisition phase.[127] Altogether, there is little evidence to support the use of concurrent feedback in distributed training. In contrast, we have recently found a large and positive effect of using summative feedback based on metrics during distributed VR simulation training: this led to a significant improvement in final-product score, metrics-based score, and number of collisions—and this increase in performance was retained after 2–3 months compared with a control group that never received summative feedback.[128]

Fidelity has for many years been a hot topic in medical education, but as discussed previously, there is little to indicate that high fidelity is better than low fidelity simulation for procedural skills training.[105] Further, the concept of fidelity and especially the dichotomisation into low and high fidelity is problematic.[106] Unfortunately, we chose to label our addition of an operating microscope eye-piece in *paper VII* as "*ultra-high fidelity*" VR simulation. Indeed, the microscope eyepiece has "ultra-high" resolution and enables a better appreciation of key visual cues in surgery compared with our standard PC screens. It should also enable improved 3D stereovision compared with the anaglyph stereo glasses that can be used with the Visible Ear Simulator. Despite the unfortunate terminology, our introduction of the

microscope eyepiece represented an attempt at increasing physical resemblance and better mimicking real life conditions where an important task is learning to view the surgical field through a microscope while handling the instruments. In contrast to our actual findings, we had expected that this would lead to an increased performance, similarly to what has been found for 3D stereovision in laparoscopy.[108]

In relation to the effects on CL, we speculated—and confirmed—that the increased realism and thereby complexity could increase the cognitive demands on the novice learner. As technology advances, we might be tempted to adopt these technological innovations in the hope that it will benefit our learners. Recently, we explored the use of head-mounted displays for immersive VR simulation of laparoscopic surgery where the simulation was virtually situated in a recorded operating room with simulated events (stressors).[129] This induced a significantly higher CL and resulted in a poorer performance compared with conventional VR simulation training.[129] This is yet another example that increasing physical resemblance and functional task alignment might not benefit learners, especially if they are novices. Even though evidence remains meagre, increased physical resemblance and functional task alignment might still have a place for more advanced learners.[105]

There is increasing interest in using patient-specific simulation in medical education but the supporting evidence is mostly limited to studies of feasibility and utility.[130] From an educational perspective, *paper VIII* adds little new: patient-specific simulation in the OSU temporal bone simulator is perceived to have potential in the training of surgeons. But in terms of streamlining the image processing and minimizing manual steps while producing accurate segmentation of temporal bone structures, *paper VIII* constitutes an important step towards clinically feasible patient-specific simulation. Other groups report that processing for patient-specific VR models averages 20–30 minutes per case[70,71] and similar to our findings, some clinical scans were not even usable for VR simulation.[71] Arora et al. further report that the visualization of key structures such as the lateral semicircular canal and facial nerve was found to be insufficient.[71] Improving segmentation and rendering for patient-specific simulation was therefore warranted and our work contributes to this.

The perspective of patient-specific temporal bone simulation goes beyond that of being merely a training tool: it might be used to plan surgical interventions, to explore different surgical approaches, and to choose for example the optimal CI electrode. This requires further modelling in the simulation data as clinical imaging modalities for example do not allow the scala vestibuli and scala tympani of the cochlea to be discerned. Augmenting the OSU temporal bone atlas[73] with the OpenEar datasets,[11] we demonstrated that the cochlear microstructures can be modelled in clinical CBCT imaging datasets.[76] We further used this modelling in pre-operative CBCT scans to determine CI electrode position in post-operative imaging.[131] This has similarly been accomplished by a group at Vanderbilt using statistical shape modelling.[72] However, the untapped potential of such modelling combine it with pre-operative, patient-specific VR simulation to plan and guide surgery. This might lead to improved patient outcomes[132] because the CI electrode position cannot be changed after insertion, and consequently the first CI insertion has to be the best possible. Ultimately, patient-specific simulation is an innovation that might prove useful not only for training but also for clinical applications.

7.3 Limitations

The limitations specific to each study have been discussed in the papers and in the following some more general limitations will briefly be discussed.

Solid empirical knowledge on the effects of different learning conditions and strategies require multiple studies[40] with interventional methodology, relevant comparisons, valid and reliable assessment outcomes including different outcome measures such as performance, learning curves, and skills transfer. This might entail randomized, clinical trials (RCTs) but these are not always the most appropriate study design in medical education.[40] We have therefore used a variety of different study designs to balance what was possible within the

confinements of setting, recruitment, time and economy, and research objective.

Most of the included studies compare different groups of learners exposed to learning interventions for example distributed practice or simulator-integrated tutoring. In most of these papers, we have termed the different groups of learners "cohorts" as they were defined by their participation in a specific learning event or condition rather than randomized for intervention or control; and the studies have accordingly been termed "educational cohort study". Nonetheless, this terminology can be challenged as it can also be argued that the studies are experimental with prospective collection of data rather than a pure observational design as cohort study could imply.

In some of the studies, data on multiple outcomes were collected and analysed separately with the lens of a specific medical educational theoretical framework or concept. Advancing and refining different conditions of learning requires deeper insights into the different effects in order to understand mechanisms and modes of action through a combination of methodological approaches. This combined data collection has not been clearly acknowledged in the individual papers and it can therefore be questioned whether for example measurement of CL through the dual-task methodology could have affected the primary performance outcome. However, the serial measurement of reaction time during simulation takes 30–60 seconds which constitutes a very limited interaction in the 30–90 minutes used for the total mastoidectomy procedure and this was not conceived or observed to impact final-product performance as it did not require the participant to stop drilling.

The included studies have primarily relied on quantitative outcome measures in controlled study designs. The controlled setting can serve as an educational laboratory where we can define the parameters and specify the learning conditions and thereby increase the signal-to-noise ratio. This facilitates interpretation of results and allows us to study in detail the effects of learning conditions in the simulation environment. Nevertheless, this is at the cost of generalizability to other contexts[133] and also presents a major challenge—namely achieving similar results when implementing learning in the complex reality of real-life settings, where many other variables might modify effects.

In several of the studies in this thesis, medical students were used to represent novices because of the very limited number of residents available. This could also have potential implications for the generalizability of the findings because effects on more advanced trainees such as residents and fellows might be very different especially for complex, subspecialized procedures such as mastoidectomy where prior knowledge, experience, and motivation plays a large role.

Next, our results highlight that there is still room for improvement of VR simulation-based training of temporal bone surgery under DSRL conditions. The literature provides a number of instructional design features that are associated with effective simulation-based learning[44] and more of these should be explored in the context temporal bone surgical training. Increasing what can be learned with simulation-based methods and minimizing loss to transfer processes can potentially further reduce the costs of subsequent clinical training. This includes reducing the amount of practice needed in the clinic and increasing patient outcomes including patient safety because the initial phases of learning, where errors are more likely, takes place in the simulation environment. For a number of reasons, the causal relationship between skills and patient-reported or clinical outcomes can be difficult to establish.[134] In temporal bone surgery, this is even more difficult because of the small number of residents that are recruited into surgical otology and the number of years from initial training during residency to independent practice. This is one of the main reasons we have used the proxy of cadaveric dissection performance to estimate effects on transfer despite this only reflecting context transfer.[47]

Learning curves are useful and yield deeper insights than end-of-training results alone.[41] Often, learning curves are reported with immediate performance measures as the learning outcome and should therefore more precisely be termed performance curves.[135] In the case of distributed practice, an element of retention is introduced into the resulting curve and this arguably brings it closer

to representing an actual learning curve, which is the reason we have termed the performance curves presented in our studies as learning curves. Regardless of terminology, these curves can be used to optimize instructional design and training programs and because both clinical and simulation training is expensive, we naturally want to achieve high-quality training but minimize the time needed for training and optimize the outcomes of training. The costs of SBT include the time the trainee uses, the time other people use (technical support, feedback and assessment, administrative support), equipment and maintenance costs, and loss of production costs if the trainee's time is dedicated to SBT over clinical duties. The costs of clinical training are even more complex and also involves human costs related to errors.

There is, however, an important consideration in relation to the use of learning curves to optimize training programs: even though the learning curve seems to plateau and the rate of learning slows down, the learning potential is not necessarily exhausted. A substantial amount of further practice might be required for true mastery. Furthermore, there might be improvement in other outcomes of learning from such "overlearning",[136] reflected for example in increasing automaticity[137] and as we demonstrated in *paper V*, a reduction of CL. Ultimately, a single-minded focus on accelerating performance during training can lead to poorer long term outcomes (the performance-learning paradox).[138] Consequently, SBT of surgical skills should balance efficiency of training and also ensure that the learning potential of SBT is close to exhausted, which is the pivot of this thesis.

A final discussion is on the limitations of expertise as a guide for training and training outcomes. In the framework of mastery learning,[22] the individual trainee must practice until proficiency at a predefined level. This level can for example be determined by the expert performance approach.[79] However, as we demonstrated, most performances of the experienced surgeons did not achieve the highest possible score. There are several potential explanations for this such as the experienced surgeons' inexperience with VR simulation, that important visual or haptic cues are missing in the simulation, or that real life is more forgiving and minor injuries are not seen in real tissues in contrast to the VR simulation.

Novice and expert performance comparisons such as those we used to develop our metrics-based score and establish proficiency levels can be problematic if cause and effect relationships are not considered.[86] Insights into how expertise is developed such as through deliberate practice[139] is not only useful for the designing training for novices because as the learner advances, other mechanisms come into play resulting for example in the "expertise reversal" effect.[60] This needs to be considered as SBT in the future is expanded to involve higher complexity learning environments (i.e., with increased physical resemblance and functional task alignment) and advanced procedures, targeting learners that are beyond the novice level. With patient-specific simulation of temporal bone surgery and simulation of advanced procedures such as CI surgery potentially moving into the clinical realm, considering the use of SBT for experienced learners is increasingly relevant.

The work presented in this thesis adds several important pieces to the current status of VR SBT and assessment of mastoidectomy based on contemporary medical educational frameworks. In most studies, our findings corroborate what is expected from theory and studies on other procedure. So why would something different be expected? Mastoidectomy is a highly complex procedure, that requires spatial understanding of important anatomy and fine motor skills. The complexity of the mastoidectomy procedure is even higher in SBT using cadavers and in real-life compared with VR simulation, and much research on CL and motor skills learning comes from other domains or studies simpler tasks. Generalizability is a double-edged sword where generalizability from one context to another might be desired but also limited. Hopefully, our results on a single procedure also translate to SBT of other complex surgical procedures despite the limitations related to study design and methodology, participants, outcomes, learning task and conditions.

7.4 Perspectives and future directions

Many of the perspectives of SBT in temporal bone surgery and directions for future research has been hinted in this thesis already.

First and foremost, self-directed VR simulation training should be the gold standard before further refinement of skills on human cadaveric temporal bones. The potential of VR simulation training far surpasses the use of VR simulation as a brief introduction immediately before dissection. VR simulation training should be systematic and structured around proficiency-based training, which can only feasibly be achieved through automated assessment for summative feedback. Further research will be needed to define levels for commencing supervised surgery and for simulation-based certification.

With the implementation of automated assessment based on our metrics-based score and with supporting tools to ensure learning a safe performance, it would be feasible to require trainees to achieve proficiency in the simulator before participation in cadaveric dissection courses. The VR simulator setup is relatively inexpensive, requires little maintenance, and is favored by trainees, which altogether makes decentralized and local training feasible. Consequently, any training department should make VR temporal bone simulation available to trainees. This would truly allow the systematic integration of VR simulation training into the curriculum. Nonetheless, implementing extensive distributed SBT into the surgical curriculum remains a challenge despite the evidence of its efficacy and needs to be addressed by educational stakeholders including program directors.

For the training of novices, there is still a need to develop and investigate other learning supports for DSRL. This could for example be to provide a variety of cases for training and ensuring motivation by introducing elements and strategies from gaming that could motivate the trainee. Tailoring feedback, instructions and case difficulty based on the individual trainees' performance and progression monitored by automated assessment would allow adaptive training programs.[49]

Increasing physical resemblance and functional task alignment is challenging but there is a need to further bridge the gap between SBT and real-life conditions and this is another important direction of research and future development relevant to advanced trainees, who have different learning needs, experiences and goals compared with novices.

Finally, clinical use of temporal bone simulation for patient-specific simulation based on clinical imaging might prove valuable for rehearsal of surgery ahead of time, for surgical planning and intervention, and for precise intra-operative navigation to further improve patient outcomes.

8. Conclusion

What we do outside, inside and in-between VR simulation matters: from valid and reliable assessment over instructional design and learning conditions to organization of practice and bringing simulation into the clinical setting.

The main conclusions that can be drawn from the studies included in this thesis are:

1. Metrics-based assessment might have a role for automated summative feedback and proficiency-based training but needs to be used in conjunction with other mechanisms to ensure that a safe performance is also learned.
2. Distributed VR simulation practice increases performance and reduces CL and acquired technical skills transfer to the cadaveric dissection environment where CL is found to be higher for novices.
3. Simulator-integrated tutoring and the use of operating microscope eyepieces during VR simulation are unfavorable for learning.

Altogether, the optimal simulation-based surgical skills training program supports DSRL with distributed practice to facilitate cognitive processes of learning, motivate the trainee, and provide direction and balanced feedback.

Future work will need to investigate how such distributed practice can be integrated into the training curriculum for example through decentralized training; how other training variables such as case variability affects performance, transfer and cognitive processes; and how feedback and assessment can be integrated for mastery learning. This could lead to optimized, adaptive training programs that automatically tailor the VR simulation training program to the individual trainee's needs based on performance, progression, and training objectives.

Ultimately, VR simulation will be useful beyond novice training and offer a valuable platform for tailoring surgery to the individual patient through patient-specific rehearsal and surgical planning in the simulation environment ahead of surgery.

9. References

1. Reznick RK. Teaching and testing technical skills. *Am J Surg* 1993;165:358–61.
2. Mills R, Lee P. Surgical skills training in middle-ear surgery. *J Laryngol Otol* 2003;117:159–63.
3. Frithioff A, Sørensen MS, Andersen SAW. European status on temporal bone training: a questionnaire study. *Eur Arch Otorhinolaryngol* 2018;275:357–63.
4. Kuppersmith RB, Johnston R, Moreau D, Loftin RB, Jenkins H. Building a virtual reality temporal bone dissection simulator. *Stud Health Technol Inform* 1997;39:180–6.
5. Mason TP, Applebaum EL, Rasmussen M, Millman A, Evenhouse R, Panko W. The virtual temporal bone. *Stud Health Technol Inform* 1998;50:346–52.
6. Agus M, Giachetti A, Gobbetti E, Zanetti G, Zorcolo A, John NW, et al. Mastoidectomy simulation with combined visual and haptic feedback. *Stud Health Technol Inform* 2002;85:17–23.
7. Pflesser B, Petersik A, Tiede U, Höhne KH, Leuwer R. Volume cutting for virtual petrous bone surgery. *Comput Aided Surg* 2002;7:74–83.
8. Morris D, Sewell C, Barbagli F, Salisbury K, Blevins NH, Girod S. Visuohaptic simulation of bone surgery for training and evaluation. *IEEE Comput Graph Appl* 2006;26:48–57.
9. Trier P, Noe KØ, Sørensen MS, Mosegaard J. The visible ear surgery simulator. *Stud Health Technol Inform* 2008;132:523–5.
10. Sørensen MS, Dobrzeniecki AB, Larsen P, Frisch T, Sporring J, Darvann TA. The visible ear: a digital image library of the temporal bone. *ORL J Otorhinolaryngol Relat Spec* 2002;64:378–81.
11. Sieber D, Erfurt P, John S, Santos GRD, Schurzig D, Sørensen MS, et al. The OpenEar library of 3D models of the human temporal bone based on computed tomography and micro-slicing. *Sci Data* 2019;6:180297.
12. Sieber DM, Andersen SAW, Sørensen MS, Mikkelsen PT. OpenEar Image Data Enables Case Variation in High Fidelity Virtual Reality Ear Surgery. *Otol Neurotol* 2021;42:1245–52.
13. The Visible Ear Simulator - academic freeware for temporal bone surgical training. Available from https://ves.alexandra.dk. [Last accessed 15/4/2022]
14. Sorensen 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.
15. Andersen SAW. Virtual reality simulation training of mastoidectomy - studies on novice performance. Dan Med J 2016;63:B5277. [Thesis]
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;141:913–8.
17. Nash R, Sykes R, Majithia A, Arora A, Singh A, Khemani S. Objective assessment of learning curves for the Voxel-Man TempoSurg temporal bone surgery computer simulator. *J Laryngol Otol* 2012;126:663–9.
18. Jowett N, LeBlanc V, Xeroulis G, MacRae H, Dubrowski A. Surgical skill acquisition with self-directed practice using computer-based video training. *Am J Surg* 2007;193:237–42.
19. Wright AS, McKenzie J, Tsigonis A, Jensen AR, Figueredo EJ, Kim S, et al. A structured self-directed basic skills curriculum results in improved technical performance in the absence of expert faculty teaching. *Surgery* 2012;151:808–14.
20. Dreyfus S, Dreyfus H. A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition. California Univ Berkeley Operations Research Center; 1980.
21. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, 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.

22. Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Mastery Learning for Health Professionals Using Technology-Enhanced Simulation: A Systematic Review and Meta-Analysis. *Acad Med* 2013;88:1178–86.
23. McGaghie WC, Issenberg SB, Barsuk JH, Wayne DB. A critical review of simulation-based mastery learning with translational outcomes. *Med Educ* 2014;48:375–85.
24. Grantcharov TP, Funch-Jensen P. Can everyone achieve proficiency with the laparoscopic technique? Learning curve patterns in technical skills acquisition. *Am J Surg* 2009;197:447–9.
25. 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.
26. Brydges R, Carnahan H, Safir O, Dubrowski A. How effective is self-guided learning of clinical technical skills? It's all about process. *Med Educ* 2009;43:507–15.
27. Brydges R, Manzone J, Shanks D, Hatala R, Hamstra SJ, Zendejas B, et al. Self-regulated learning in simulation-based training: a systematic review and meta-analysis. *Med Educ* 2015;49:368–78.
28. Devine LA, Donkers J, Brydges R, Perelman V, Cavalcanti RB, Issenberg SB. An Equivalence Trial Comparing Instructor-Regulated With Directed Self-Regulated Mastery Learning of Advanced Cardiac Life Support Skills. *Simul Healthc* 2015;10:202–9.
29. 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.
30. Carraccio C, Wolfsthal SD, Englander R, Ferentz K, Martin C. Shifting paradigms: from Flexner to competencies. *Acad Med* 2002;77:361–7.
31. Martin JA, Regehr G, Reznick R, Macrae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997;84:273–8.
32. 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. *Acad Med* 2013;88:872–83.
33. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006;119:166.e7-16.
34. Zirkle M, Taplin MA, Anthony R, Dubrowski A. Objective assessment of temporal bone drilling skills. *Ann Otol Rhinol Laryngol* 2007;116:793–8.
35. Laeeq K, Bhatti NI, Carey JP, Della Santina CC, Limb CJ, Niparko JK, et al. Pilot testing of an assessment tool for competency in mastoidectomy. *Laryngoscope* 2009;119:2402–10.
36. Francis HW, Masood H, Chaudhry KN, Laeeq K, Carey JP, Della Santina CC, et al. Objective assessment of mastoidectomy skills in the operating room. *Otol Neurotol* 2010;31:759–65.
37. Butler NN, Wiet GJ. Reliability of the Welling scale (WS1) for rating temporal bone dissection performance. *Laryngoscope* 2007;117:1803–8.
38. Sethia R, Kerwin TF, Wiet GJ. Performance Assessment for Mastoidectomy. *Otolaryngol Head Neck Surg* 2017;156:61–9.
39. Satava RM, Cuschieri A, Hamdorf J, et al. Metrics for objective Assessment. *Surg Endosc* 2003;17:220–6.
40. Norman G. RCT = results confounded and trivial: the perils of grand educational experiments. *Med Educ* 2003;37:582–4.
41. Pusic MV, Boutis K, Hatala R, Cook DA. Learning Curves in Health Professions Education. *Acad Med* 2015;90:1034–42.
42. Magill R. Motor Learning and Control: Concepts and Applications. New York, NY: McGraw-Hills; 2013.
43. Konge L, Clementsen PF, Ringsted C, Minddal V, Larsen KR, Annema JT. Simulator training for endobronchial ultrasound: a randomised controlled trial. *Eur Respir J* 2015;46:1140–9.
44. Cook DA, Hamstra SJ, Brydges R, Zendejas B, Szostek JH, Wang AT, et al. Comparative effectiveness of instructional design features in simulation-based education: systematic review and meta-analysis. *Med Teach* 2013;35:e867-898.
45. Eva KW, Neville AJ, Norman GR. Exploring the etiology of content specificity: factors influencing analogic transfer and problem solving. *Acad Med* 1998;73:S1-5.
46. Tolsgaard MG. Assessment and learning of ultrasound skills in Obstetrics & Gynecology. *Dan Med J* 2018;65:B5445. [Thesis]
47. Haskell R. Transfer of learning: What it is and why it's important, in "Transfer of Learning: Cognition and Instruction". Academic Press; 2001.
48. Sweller J, Ayres P, Kalyuga S. Cognitive Load Theory. New York: Springer; 2011.
49. van Merriënboer JJG, Sweller J. Cognitive Load Theory and Complex Learning: Recent Developments and Future Directions. *Educ Psychol Rev* 2005;17:147–77.
50. Sweller J, van Merriënboer JJG, Paas F. Cognitive architecture and instructional design: 20 years later. *Educ Psychol Rev* 2019;31:261–92.
51. van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: design principles and strategies. *Med Educ* 2010;44:85–93.
52. Brünken R, Plass JL, Leutner D. Direct measurement of cognitive load in multimedia learning. *Educ Psychol* 2003;38:53–61.
53. Naismith LM, Cavalcanti RB. Validity of Cognitive Load Measures in Simulation-Based Training: A Systematic Review. *Acad Med* 2015;90:S24-35.
54. Park B, Brünken R. The Rhythm Method: A New Method for Measuring Cognitive Load—An Experimental Dual-Task Study. *Appl Cogn Psychol* 2015;29:232–43.
55. Cao CGL, Zhou M, Jones DB, Schwaitzberg SD. Can surgeons think and operate with haptics at the same time? *J Gastrointest Surg* 2007;11:1564–9.
56. Haji FA, Rojas D, Childs R, de Ribaupierre S, Dubrowski A. Measuring cognitive load: performance, mental effort and simulation task complexity. *Med Educ* 2015;49:815–27.
57. Rojas D, Haji F, Shewaga R, Kapralos B, Dubrowski A. The impact of secondary-task type on the sensitivity of reaction-time based measurement of cognitive load for novices learning surgical skills using simulation. *Stud Health Technol Inform* 2014;196:353–9.
58. Bharathan R, Vali S, Setchell T, Miskry T, Darzi A, Aggarwal R. Psychomotor skills and cognitive load training on a virtual reality laparoscopic simulator for tubal surgery is effective. *Eur J Obstet Gynecol Reprod Biol* 2013;169:347–52.
59. Wulf G, Shea CH. Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychon Bull Rev* 2002;9:185–211.
60. van Gog T, Ericsson KA, Rikers RMJP, Paas F. Instructional Design for Advanced Learners: Establishing Connections Between the Theoretical Frameworks of Cognitive Load and Deliberate Practice. *Educ Technol Res Dev* 2005;53:73–81.
61. Swartz JD, Loevner LA. Imaging of the Temporal Bone. Thieme; 2009.
62. Miracle AC, Mukherji SK. Conebeam CT of the Head and Neck, Part 1: Physical Principles. *Am J Neuroradiol* 2009;30:1088–95.
63. Miracle AC, Mukherji SK. Conebeam CT of the Head and Neck, Part 2: Clinical Applications. *Am J Neuroradiol* 2009;30:1285–92.
64. Güldner C, Diogo I, Bernd E, Dräger S, Mandapathil M, Teymoortash A, et al. Visualization of anatomy in normal and pathologic middle ears by cone beam CT. *Eur Arch Otorhinolaryngol* 2017;274:737–42.
65. Peltonen LI, Aarnisalo AA, Kortesniemi MK, Suomalainen A, Jero J, Robinson S. Limited cone-beam computed tomography imaging of the middle ear: a comparison with multislice helical computed tomography. *Acta Radiol* 2007;48:207–12.
66. Dahmani-Causse M, Marx M, Deguine O, Fraysse B, Lepage B, Escudé B. Morphologic examination of the temporal bone by cone beam computed tomography: Comparison with multislice helical computed tomography. *Eur Ann Otorhinolaryngol Head Neck Dis* 2011;128:230–5.
67. Otoplan Otological Software, CAScination, Bern, Switzerland. https://otoplan.ch [Last accessed 15/4/2022].

68. Sethia R, Wiet GJ. Preoperative preparation for otologic surgery: temporal bone simulation. *Curr Opin Otolaryngol Head Neck Surg* 2015;23:355–9.
69. Andersen SAW, Bergman M, Keith JP, Powell KA, Hittle B, Malhotra P, et al. Segmentation of Temporal Bone Anatomy for Patient-Specific Virtual Reality Simulation. *Ann Otol Rhinol Laryngol* 2020:000348942097021.
70. Chan S, Li P, Locketz G, Salisbury K, Blevins NH. High-fidelity haptic and visual rendering for patient-specific simulation of temporal bone surgery. *Comput Assist Surg (Abingdon)* 2016;21:85–101.
71. Arora A, Swords C, Khemani S, Awad Z, Darzi A, Singh A, et al. Virtual reality case-specific rehearsal in temporal bone surgery: A preliminary evaluation. *Int J Surg* 2014;12:141–5.
72. Noble JH, Gifford RH, Labadie RF, Dawant BM. Statistical shape model segmentation and frequency mapping of cochlear implant stimulation targets in CT. *Med Image Comput Comput Assist Interv* 2012;15:421–8.
73. Powell KA, Liang T, Hittle B, Stredney D, Kerwin T, Wiet GJ. Atlas-Based Segmentation of Temporal Bone Anatomy. *Int J Comput Assist Radiol Surg* 2017;12:1937–44.
74. Powell KA, Kashikar T, Hittle B, Stredney D, Kerwin T, Wiet GJ. Atlas-based segmentation of temporal bone surface structures. *Int J Comput Assist Radiol Surg* 2019;14:1267–73.
75. Wang J, Lv Y, Wang J, Ma F, Du Y, Fan X, et al. Fully automated segmentation in temporal bone CT with neural network: a preliminary assessment study. *BMC Med Imaging* 2021;21:166.
76. Powell KA, Wiet GJ, Hittle B, Oswald GI, Keith JP, Stredney D, et al. Atlas-based segmentation of cochlear microstructures in cone beam CT. *Int J Comput Assist Radiol Surg* 2021;16:363–373.
77. Andersen SAW, Cayé-Thomasen P, Sørensen MS. Mastoidectomy performance assessment of virtual simulation training using final-product analysis. *Laryngoscope* 2015;125:431–5.
78. Al-Shahrestani F, Sørensen MS, Andersen SAW. Performance metrics in mastoidectomy training: a systematic review. *Eur Arch Otorhinolaryngol* 2019;276:657–64.
79. Causer J, Barach P, Williams AM. Expertise in medicine: using the expert performance approach to improve simulation training. *Med Educ* 2014;48:115–23.
80. Sewell C, Morris D, Blevins NH, Barbagli F, Salisbury K. Evaluating drilling and suctioning technique in a mastoidectomy simulator. *Stud Health Technol Inform* 2007;125:427–32.
81. Ioannou I, Avery A, Zhou Y, Szudek J, Kennedy G, O'Leary S. The effect of fidelity: how expert behavior changes in a virtual reality environment. *Laryngoscope* 2014;124:2144–50.
82. Ioannou I, Zhou Y, Wijewickrema S, Piromchai P, Copson B, Kennedy G, et al. Comparison of Experts and Residents Performing a Complex Procedure in a Temporal Bone Surgery Simulator. *Otol Neurotol* 2017;38:e85–91.
83. Wan D, Wiet GJ, Welling DB, Kerwin T, Stredney D. Creating a cross-institutional grading scale for temporal bone dissection. *Laryngoscope* 2010;120:1422–7.
84. Kerwin T, Wiet G, Stredney D, Shen H-W. Automatic scoring of virtual mastoidectomies using expert examples. *Int J Comput Assist Radiol Surg* 2012;7:1–11.
85. Kerwin T, Stredney D, Wiet G, Shen H-W. Virtual mastoidectomy performance evaluation through multi-volume analysis. *Int J Comput Assist Radiol Surg* 2013;8:51–61.
86. Cook DA. Much ado about differences: why expert-novice comparisons add little to the validity argument. *Adv Health Sci Educ Theory Pract* 2015;20:829–34.
87. Bloch R, Norman G. Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. *Med Teach* 2012;34:960–92.
88. Brennan RL. Generalizability theory and classical test theory. *Appl Meas Educ* 2011;24:1–21.
89. Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ* 2004;38:1006–12.
90. Fernandez SA, Wiet GJ, Butler NN, Welling B, Jarjoura D. Reliability of surgical skills scores in otolaryngology residents: analysis using generalizability theory. *Eval Health Prof* 2008;31:419–36.
91. Andersen SAW, Park YS, Sørensen MS, Konge L. Reliable Assessment of Surgical Technical Skills Is Dependent on Context: An Exploration of Different Variables Using Generalizability Theory. *Acad Med* 2020;95:1929–36.
92. Moulton C-AE, Dubrowski A, Macrae H, Graham B, Grober E, Reznick R. Teaching surgical skills: what kind of practice makes perfect?: a randomized, controlled trial. *Ann Surg* 2006;244:400–9.
93. Mackay S, Morgan P, Datta V, Chang A, Darzi A. Practice distribution in procedural skills training: a randomized controlled trial. *Surg Endosc* 2002;16:957–61.
94. Andersen SAW, Konge L, Cayé-Thomasen P, Sørensen MS. Retention of Mastoidectomy Skills After Virtual Reality Simulation Training. *JAMA Otolaryngol Head Neck Surg* 2016;142:635–40.
95. Malekzadeh S, Malloy KM, Chu EE, Tompkins J, Battista A, Deutsch ES. ORL emergencies boot camp: using simulation to onboard residents. *Laryngoscope* 2011;121:2114–21.
96. Chin CJ, Chin CA, Roth K, Rotenberg BW, Fung K. Simulation-based otolaryngology - head and neck surgery boot camp: "how I do it." *J Laryngol Otol* 2016;130:284–90.
97. 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.
98. Frendø M, Thingaard E, Konge L, Sørensen MS, Andersen SAW. Decentralized virtual reality mastoidectomy simulation training: a prospective, mixed-methods study. *Eur Arch Otorhinolaryngol* 2019;276:2783–9.
99. 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.
100. Aldekhyl S, Cavalcanti RB, Naismith LM. Cognitive load predicts point-of-care ultrasound simulator performance. *Perspect Med Educ* 2018;7:23–32.
101. Rasmussen SR, Konge L, Mikkelsen PT, Sørensen MS, Andersen SAW. Notes From the Field: Secondary Task Precision for Cognitive Load Estimation During Virtual Reality Surgical Simulation Training. *Eval Health Prof* 2016;39:114–20.
102. Shea CH, Lai Q, Black C, Park J-H. Spacing practice sessions across days benefits the learning of motor skills. *Hum Mov Sci* 2000;19:737–60.
103. Hatala R, Cook DA, Zendejas B, Hamstra SJ, Brydges R. Feedback for simulation-based procedural skills training: a meta-analysis and critical narrative synthesis. *Adv Health Sci Educ Theory Pract* 2014;19:251–72.
104. Andersen S, Cayé-Thomasen P, Sørensen M. Novices perform better in virtual reality simulation than in traditional cadaveric dissection training of mastoidectomy. *J Surg Simul* 2015;2:68–75.
105. Lefor AK, Harada K, Kawahira H, Mitsuishi M. The effect of simulator fidelity on procedure skill training: a literature review. *Int J Med Educ* 2020;11:97–106.
106. Hamstra SJ, Brydges R, Hatala R, Zendejas B, Cook DA. Reconsidering Fidelity in Simulation-Based Training. *Acad Med* 2014;89:6.
107. Rangarajan K, Davis H, Pucher PH. Systematic Review of Virtual Haptics in Surgical Simulation: A Valid Educational Tool? *J Surg Educ* 2020;77:11.
108. Sørensen SMD, Savran MM, Konge L, Bjerrum F. Three-dimensional versus two-dimensional vision in laparoscopy: a systematic review. *Surg Endosc* 2016;30:11–23.
109. Aggarwal R, Brown KM, Gallagher AG, Henriksen K, Peng GCY, Ritter EM, et al. Simulation Research in Gastrointestinal and Urologic Care—Challenges and Opportunities. *Ann Surg* 2017:9.
110. Riojas KE, Tran ET, Freeman MH, Noble JH, Webster 3rd RJ, Labadie RF. Clinical Translation of an Insertion Tool for

Minimally Invasive Cochlear Implant Surgery. *J Med Devices* 2021;15:031001.

111. Frithioff A, Frendø M, Pedersen DB, Sørensen MS, Wuyts Andersen SA. 3D-Printed Models for Temporal Bone Surgical Training: A Systematic Review. *Otolaryngol Head Neck Surg* 2021;165:617–25.
112. Locketz GD, Lui JT, Chan S, Salisbury K, Dort JC, Youngblood P, et al. Anatomy-Specific Virtual Reality Simulation in Temporal Bone Dissection: Perceived Utility and Impact on Surgeon Confidence. *Otolaryngol Neck Surg* 2017;156:1142–9.
113. Lui J, Locketz G, Dort J, Chen J, Chau J, Chan S, et al. Assessing round window depiction in a virtual reality environment for cochlear implantation. *J Surg Simul* 2021;7:85–91.
114. Andersen SAW, Nayahangan LJ, Park YS, Konge L. Use of Generalizability Theory for Exploring Reliability of and Sources of Variance in Assessment of Technical Skills: A Systematic Review and Meta-Analysis. *Acad Med* 2021;96:1609–19.
115. Preisler L, Søndergaard Svendsen MB, Søndergaard B, Brink L, Nordentoft T, Svendsen LB, et al. Automatic and unbiased assessment of competence in colonoscopy: exploring validity of the Colonoscopy Progression Score (CoPS). *Endosc Int Open* 2016;4:E1238–43.
116. Sewell C, Morris D, Blevins N, Barbagli F, Salisbury K. Quantifying risky behavior in surgical simulation. *Stud Health Technol Inform* 2005;111:451–7.
117. Yudkowsky R, Park YS, Downing SM. Assessment in Health Professions Education. Routledge; 2019.
118. Pietersen PI, Bjerrum F, Tolsgaard MG, Konge L, Andersen SAW. Standard Setting in Simulation-Based Training of Surgical Procedures: A Systematic Review. *Ann Surg* 2021 Sep 13 [Epub ahead of print].
119. Gustafsson A, Pedersen P, Rømer TB, Viberg B, Palm H, Konge L. Hip-fracture osteosynthesis training: exploring learning curves and setting proficiency standards. *Acta Orthop* 2019;90:348–53.
120. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. Effect of practice on standardised learning outcomes in simulation-based medical education. *Med Educ* 2006;40:792–7.
121. Lui JT, Hoy MY. Evaluating the Effect of Virtual Reality Temporal Bone Simulation on Mastoidectomy Performance: A Meta-analysis. *Otolaryngol Neck Surg* 2017;156:1018–24.
122. Gawęcki W, Węgrzyniak M, Mickiewicz P, Gawłowska MB, Talar M, Wierzbicka M. The Impact of Virtual Reality Training on the Quality of Real Antromastoidectomy Performance. *J Clin Med* 2020;9:3197.
123. Andersen SAW, Konge L, Sørensen MS. The effect of distributed virtual reality simulation training on cognitive load during subsequent dissection training. *Med Teach* 2018;40:684–9.
124. Schnotz W, Kürschner C. A reconsideration of cognitive load theory. *Educ Psychol Rev* 2007;19:469–508.
125. Klepsch M, Schmitz F, Seufert T. Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Front Psychol* 2017;8 :1997.
126. Wijewickrema S, Piromchai P, Zhou Y, Ioannou I, Bailey J, Kennedy G, et al. Developing Effective Automated Feedback in Temporal Bone Surgery Simulation. *Otolaryngol Neck Surg* 2015;152:1082–8.
127. Schmidt RA, Wulf G. Continuous concurrent feedback degrades skill learning: implications for training and simulation. *Hum Factors* 1997;39:509–25.
128. 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.
129. Frederiksen JG, Sørensen SMD, Konge L, Svendsen MBS, Nobel-Jørgensen M, Bjerrum F, et al. Cognitive load and performance in immersive virtual reality versus conventional virtual reality simulation training of laparoscopic surgery: a randomized trial. *Surg Endosc* 2020;34:1244–52.
130. Ryu WHA, Dharampal N, Mostafa AE, Sharlin E, Kopp G, Jacobs WB, et al. Systematic Review of Patient-Specific Surgical Simulation: Toward Advancing Medical Education. *J Surg Educ* 2017;74:1028–38.
131. Andersen SAW, Keith JP, Hittle B, Riggs WJ, Adunka O, Wiet GJ, et al. Automated Calculation of Cochlear Implant Electrode Insertion Parameters in Clinical Cone-Beam CT. *Otol Neurotol* 2022;43:199–205.
132. Chakravorti S, Noble JH. Further Evidence of the Relationship Between Cochlear Implant Electrode Positioning and Hearing Outcomes. *Otol Neurotol* 2019;40:617-624.
133. Eva KW. Broadening the debate about quality in medical education research. *Med Educ* 2009;43:294–6.
134. Cook DA, West CP. Perspective: Reconsidering the focus on "outcomes research" in medical education: a cautionary note. *Acad Med* 2013;88:162–7.
135. Dubrowski A. Performance vs. learning curves: what is motor learning and how is it measured? *Surg Endosc* 2005;19:1290.
136. Kolozsvari NO, Kaneva P, Brace C, Chartrand G, Vaillancourt M, Cao J, et al. Mastery versus the standard proficiency target for basic laparoscopic skill training: effect on skill transfer and retention. *Surg Endosc* 2011;25:2063–70.
137. Stefanidis D, Scerbo MW, Sechrist C, Mostafavi A, Heniford BT. Do novices display automaticity during simulator training? *Am J Surg* 2008;195:210–3.
138. Schmidt RA, Bjork RA. New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychol Sci* 1992;3:207–17.
139. Ericsson KA. Deliberate Practice and the Acquisition and Maintenance of Expert Performance in Medicine and Related Domains. *Acad Med* 2004;79:S70–81.

ISBN: 978-8-79748-720-4