UNIVERSITY OF COPENHAGEN
FACULTY OF HEALTH AND MEDICAL SCIENCES

# PhD thesis

Martin Frendø, MD

## Virtual Reality simulation-based training of cochlear implant surgery: perspectives of performance, assessment, and transfer

Principal supervisor: Professor Mads Sølvsten Sørensen, MD, DMSc.

This thesis has been submitted to The Graduate School of Health and Medical Sciences, University of Copenhagen, February 25, 2021

# Virtual Reality simulation-based training of cochlear implant surgery: perspectives of performance, assessment, and transfer

PhD thesis by Martin Frendø, MD

Department of Otorhinolaryngology—Head & Neck Surgery and Audiology
Copenhagen University Hospital, Rigshospitalet
Denmark

**Academic supervisors**

**Principal supervisor:**
Professor Mads Sølvsten Sørensen, MD, DMSc.
Department of Otorhinolaryngology—Head & Neck Surgery and Audiology
Copenhagen University Hospital, Rigshospitalet

**Primary co-supervisor:**
Professor Lars Konge, MD, PhD
Copenhagen Academy for Medical Education and Simulation, Center for HR & Education
Copenhagen University Hospital, Rigshospitalet

**Co-supervisor:**
Steven Arild Wuyts Andersen, MD, PhD
Department of Otorhinolaryngology—Head & Neck Surgery and Audiology,
Copenhagen University Hospital, Rigshospitalet and
Copenhagen Academy for Medical Education and Simulation, Center for HR & Education

**Assessment committee**

Professor Morten Dornonville de la Cour, MD, DMSc., FEBO, MPG (chairman)
Department of Ophthalmology
Copenhagen University Hospital, Rigshospitalet

Professor Joachim Müller, MD, DMSc.
Department of Otorhinolaryngology, Head and Neck Surgery
University Medical Center Munich
Ludwig Maximilian University, Munich, Germany

Professor Teodor Grantcharov, MD, PhD, FACS
Department of Surgery
St. Michael's Hospital
University of Toronto, Canada

Front-page photo of cochlear implant device © Oticon Medical

## Acknowledgments

I genuinely enjoyed and had fun doing the research for this thesis. This would not have been the case without excellent guidance and supervision from my three PhD supervisors:

Professor, MD, DMSc. Mads Sølvsten Sørensen started the Visible Ear Simulator project, providing Virtual Reality (VR) temporal bone surgery simulation at a high quality and a low cost. This inspirational common good approach permeates Mads' work. Mads gave me the freedom to find my own way of doing things while always helping and giving advice when needed.

MD, PhD Steven A.W. Andersen introduced me to the wonderful world of educational science and was always generous with his time and help throughout the studies.

Professor, MD, PhD, Lars Konge had the educational research question and hypotheses at the forefront. At the Copenhagen Academy for Medical Education and Simulation, Lars created a unique and stimulating research environment.

Overall, I have been very fortunate to work with Mads, Steven and Lars and am grateful for the collaboration.

I also thank my co-authors: professor, MD, DMSc. Per Cayé-Thomasen, associate professor, MD, Søren Foghsgaard, and computer scientist, MSc Peter Trier Mikkelsen.

Next, I would like to thank both administrative and academic staff at the Copenhagen Academy for Medical Education and Simulation.

Professor Gregory J. Wiet, Ohio State University, Columbus, Ohio, USA, generously hosted my research stay—thank you.

I thank my colleagues at the Department of Otorhinolaryngology—Head and Neck Surgery & Audiology: Mads Klokker, Berit Merete Sachmann, Jacob Melchiors, Henrik Bredahl, Ramon G. Jensen, and Christian von Buchwald.

I thank the many otologists, students and residents who enabled the studies by partaking, as well as associate professor, MD, PhD Grethe Samuelsen and associate professor, MD, DMSc. Martin Nue Møller for help with dissection courses.

I also thank the Oticon Foundation who supported the department with the funds for this project, and Mette Bülow Clausen, Oticon Medical, who provided electrodes for Cochlear Implant studies.

I thank MD, PhD-fellow Andreas Frithioff with whom I collaborated during the data collection for Study I–III and Andreas' subsequent PhD studies. Few get things done like Andreas, and it has been a joy working with him. Further, I thank BSc.med. Karoline A. Arnesen and MD Josefine H. von Buchwald for collaboration on their master's theses.

I thank my MD colleagues at the CAMES PhD office. We have had so much fun together: from daily banter and countless academic discussions to sharing the inevitable ups and downs of research.

I sincerely thank my parents Lisbeth and Peter-Ove for unwavering support.

Last but certainly not least, I wholeheartedly thank Sofie and Olav for relentless support and encouragement, without which this work would not have been possible.

# Dedication

With gratitude, this thesis is dedicated to

**Sten Visti Jensen**

Cand. mag.

# Table of Contents

# Abbreviations

| | |
|---|---|
| AL | apprenticeship learning |
| CI | cochlear implant |
| CISAT | Cochlear Implant Surgery Assessment Tool |
| DSRL | directed, self-regulated learning |
| OR | operating room |
| ORL | otorhinolaryngology |
| PHLD | profound hearing loss or deafness |
| RW | round window |
| SBT | simulation-based training |
| VES | Visible Ear Simulator |
| VR | virtual reality |

## Summary in English

Cochlear implant (CI) surgery is the most effective treatment for deafness and profound hearing loss. The CI procedure requires surgical expertise because vital anatomical structures and residual hearing can be damaged. To acquire this, deliberate practice is needed but training options outside the high-risk operating room (OR) are limited, and prerequisites for deliberate practice lacking. Virtual Reality (VR) simulation-based training (SBT) is effective for mastoidectomy training, but little is known about CI VR SBT and even for mastoidectomy VR SBT, implementation is lacking.

The objectives of this PhD thesis were 1) to investigate evidence-based assessment of CI surgical skills in SBT by developing a new assessment instrument for CI VR SBT; 2) to evaluate the acquisition of skills during SBT (i.e. the learning curve of training) and transfer of CI electrode insertion skills from VR to a 3D-printed model; 3) to assess the transfer of skills from CI VR SBT to cadaver dissection; and 4) to explore a novel implementation of VR SBT of mastoidectomy: as decentralized training, where the trainee learns independently, devoid of direct hands-on supervision but with simulator-integrated learning supports.

First, we developed an assessment tool for evaluating CI surgery skills: the Cochlear Implant Surgery Assessment Tool (CISAT). We gathered validity evidence to support the assessment according to Messick's validity framework and determined a pass/fail score, which can be used for mastery learning. The CISAT was determined to be highly reliable for CI VR SBT skills assessment.

Secondly, we explored the learning curve during CI VR SBT and found that skills acquisition was highly heterogeneous, and on average followed a negatively accelerated learning curve pattern. However, even after 18 procedures, novices had not reached a learning curve plateau. This suggests that reaching a stable performance level during CI VR SBT requires a substantial training volume, and that training can lead to improvement of skills even after 18 training procedures. Transfer of insertion skills to a 3D-printed temporal bone model was modest.

In the third study, a randomized controlled trial, we piloted skills transfer of two hours of CI VR SBT to cadaveric dissection performance as an addition to the standard mastoidectomy training given the control group. We found that trainees in the intervention group performed marginally—and statistically insignificantly—better than the control group. However, the intervention group seemed to require less instructor support during cadaver dissection. Correspondingly, the pilot study did not demonstrate a substantial training effect of the CI VR SBT during cadaver dissection, either due to insufficient volume or effect of the CI VR SBT, or due to methodological problems, most notably the small study size.

The fourth study explored implementing VR simulation as decentralized training, where the participants trained at their local department or at home, aided by learning supports. We found that the intervention cohort markedly outperformed the control cohort. This suggests that decentralized VR training for directed, self-regulated learning is feasible and efficient even in the absence of dedicated training time during work hours.

Altogether, these studies demonstrate 1) that CI skills during VR SBT can be assessed validly and reliably; 2) the individuality of learning curves implies that CI training should pivot on objective assessment of performance and the lack of plateau that, unlike isolated mastoidectomy, CI VR SBT requires a lot of practice before a stable level is reached; 3) transfer of CI surgery skills to cadaver SBT seems modest, but requires further investigation due to methodological issues, most notably a low number of training procedures and participants; 4) decentralized training is feasible and effective for mastoidectomy VR SBT.

These findings can be used to develop evidence-based training curricula. This might improve training of temporal bone surgery by allowing tracking of performance during training, evaluation of training interventions, or for competence-based progression during training (mastery learning). Finally, decentralized training can provide distributed practice at the convenience of the trainee.

# Resumé på dansk

Cochlear implant (CI) kirurgi er den mest effektive behandling af døvhed og svært høretab. CI proceduren kræver kirurgisk ekspertise, da vigtige anatomiske strukturer og resthørelse ellers kan kompromitteres. Målrettet træning er påkrævet for at opnå dette, men træningsmulighederne er ofte begrænsede og forudsætningerne for målrettet træning utilstrækkelige. Virtual reality (VR) simulations-baseret træning (SBT) er en effektiv træningsmetode af mastoidektomi, men kendskabet til CI VR SBT er minimalt. Det samme gælder implementering af VR SBT af mastoidektomi.

Formålet med denne afhandling var at udarbejde og evaluere evidensbaseret kompetencevurdering ved CI-kirurgi, inklusiv CI VR SBT; at undersøge læringskurven ved CI VR SBT, samt at studere overførsel af de tillærte evner fra VR til kadavertræning. Slutteligt ønskede vi at undersøge en ny måde at træne VR SBT, nemlig som decentral træning, hvor den uddannelsessøgende selvstændigt skal træne lokalt uden direkte supervision eller instruktion.

Afhandlingen indeholder fire studier. I det første studie udviklede vi et kompetencevurderingsværktøj til CI kirurgi (CISAT) og indsamlede validitetsbeviser ved hjælp af Messick's validitetsmodel. Vi fandt at CISAT understøttedes af de indsamlede validitetsbeviser og vi fastsatte endvidere en grænse for bestået/ikke bestået som kan bruges til mestringsindlæring. Analyse af CISAT-værktøjets pålidelighed viste at det var yderst pålideligt til vurdering af færdigheder ved CI VR SBT.

I næste studie undersøgte vi læringskurven ved CI VR SBT, som viste stor individuel variation. Overordnet sås et traditionelt, negativt accelereret læringskurvemønster, dog uden et læringskurveplateau. Det antyder at CI VR SBT, modsat VR SBT af mastoidektomi, fortsat medfører læring efter 18 procedurer. Overførsel af CI insertions-evner til en 3D-printet model var beskeden, muligvis fordi den 3D-printede model ikke er understøttet af pålidelige validitetsbeviser, og ikke er tilstrækkeligt naturtro.

I tredje studie undersøgte vi overførsel af kompetencer fra VR til kadavertræning i et lodtræknings-pilotstudie. Her fik interventionsgruppen to timers CI VR SBT i tillæg til den standardtræning som blev givet kontrolgruppen. Resultaterne viste en lille og statistisk ikke-signifikant forbedring i interventionsgruppen, der dog krævede mindre hjælp undervejs i indgrebet end kontrolgruppen. Studiet formåede ikke at demonstrere en substantiel effekt af interventionen. Dette kan skyldes at interventionen simpelthen ikke virker (f.eks. grundet for lille træningsvolumen); alternativt metodiske problemer, navnligt at studiets begrænsede deltagerantal ikke var tilstrækkeligt til at vise en effekt.

I fjerde studie undersøgte vi en ny implementering af VR SBT af mastoidektomi: som decentral træning på den lokale afdeling eller hjemme. Ved at evaluere effekten af denne træning ved efterfølgende kadavertræning, fandt vi at interventionskohorten klart overgik kontrolkohorten. Det indikerer at decentral VR SBT hjulpet af tilstrækkelig læringsstøtte til vejledt, selvreguleret træning, er gennemførlig og effektiv, selv i fravær af dedikeret træningstid.

Vores fund demonstrerer at CI-kirurgiske evner kan vurderes validt og pålideligt ved CI VR SBT. Læringskurvernes store individuelle variation indikerer at træning af CI kirurgi bør have objektiv kompetencevurdering som omdrejningspunkt, da antal procedurer ikke tilstrækkeligt prædikterer den enkeltes niveau. Fraværet af et læringskurveplateau efter 18 procedurer indikerer at CI VR SBT kræver væsentligt mere træning end mastoidektomi VR SBT før et stabilt niveau nås. Overførsel af kirurgiske evner fra VR SBT til kadaver CI træning synes beskedent, men kræver yderligere undersøgelse, idet metodiske problemer (særligt den lille træningsmængde og studiets ringe statistiske styrke) kan maskere en eventuel effekt. Slutteligt udgør decentral træning en gennemførlig og effektiv interventionsmetode for VR SBT af mastoidektomi.

Afhandlingens fund kan bruges i programmer for træning af tindingebenskirurgi ved at give mulighed for monitorering af det kirurgiske niveau under oplæring, evaluering af træningsinterventioner, og til kompetence-baseret progression af træningen (mastery learning). Slutteligt kan decentral træning give lejlighed til distribueret træning på en måde, der er belejlig for den uddannelsessøgende.

# Publication overview

I. Frendø M, Frithioff A, Konge L, Foghsgaard S, Mikkelsen PT, Sørensen MS, Cayé-Thomasen P, Andersen SAW. *Assessing competence in Cochlear Implant Surgery using the newly developed Cochlear Implant Surgery Assessment Tool.* Eur Arch Otorhinolaryngol. 2021 Feb 19. doi: 10.1007/s00405-021-06632-9. Online ahead of print.

II. Frendø M, Frithioff A, Konge L, Sørensen MS, Andersen SAW. *Cochlear implant surgery: Learning curve in virtual reality simulation training and transfer of skills to a 3D-printed temporal bone—a prospective trial.* Submitted to *Cochlear Implants International*

III. Frendø M, Frithioff A, Konge L, Cayé-Thomasen P, Sørensen MS, Andersen SAW. *Cochlear implant surgery: Virtual reality simulation training and transfer of skills to cadaver dissection—a randomized, controlled pilot study.* Ready for submission.

IV. Frendø M, Konge L, Cayé-Thomasen P, Sørensen MS, Andersen SAW. *Decentralized Virtual Reality Training of Mastoidectomy Improves Cadaver Dissection Performance: A Prospective, Controlled Cohort Study.* Otol Neurotol. 2020 Apr;41(4):476-481.

# Background

## Introduction

Approximately 4‰ of the world's population is estimated to have profound hearing loss or deafness (PHLD)[1–4]—a number expected to almost double by 2050[5]. PHLD substantially affects quality of life[6], employment[7,8], and frequently leads to stigma, grief and social exclusion[9]. From a population-based standpoint, PHLD are costly conditions: healthcare costs, indirect costs (e.g. unemployment), and intangible/social costs make PHLD both individual and a societal problems[10,11]. Cochlear implantation is an effective treatment of PHLD. A cochlear implant (CI) can induce or partially restore hearing by electrically stimulating the spiral ganglion cells of the cochlea. CI surgery is considered one of the most significant medico-technical inventions of the 20th century[12]. Essential to any procedure—including cochlear implantation—is the surgical performance[13,14]. Novices are prone to making surgical errors because they lack skills and experience[15]. Surgical training aims to alleviate this, but traditional training is challenged. Simulation-based training (SBT) offers a possible solution to basic training needs before further honing of surgical skills through supervised training in the OR. In SBT such as Virtual Reality (VR) SBT, trainees can acquire basic skills before cadaver SBT and real-life surgery. Unfortunately, knowledge to inform decision-makers and trainees on SBT of CI surgery is inadequate. Although VR SBT of mastoidectomy is supported by evidence, pragmatic studies on real-life implementation are lacking. Altogether, there is a need for studies exploring CI SBT, and the resulting learning inside and outside the VR simulation environment, as well as implementation of mastoidectomy VR SBT in the training curriculum. In this thesis, we describe four studies that address this need.

## Mastoidectomy and Cochlear Implant (CI) surgery

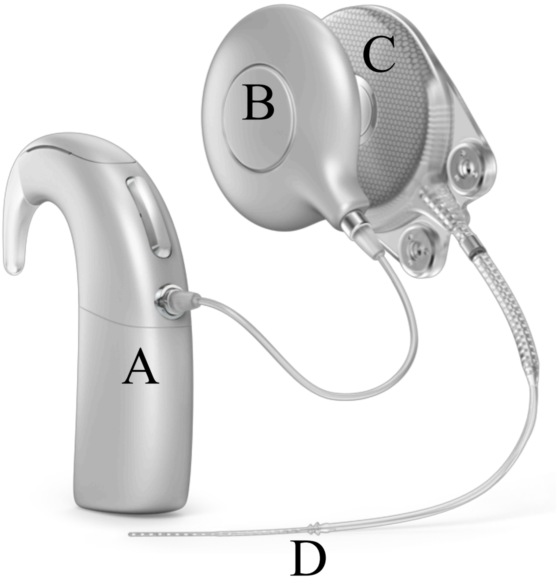
The mastoidectomy procedure comprises drilling of the mastoid in the temporal bone and is used for different purposes: to treat acute mastoiditis, remove cholesteatoma or vestibular schwannoma, or to gain access to the cochlea for cochlear implantation. Injuring adjacent structures can cause serious adverse events: damage to the facial nerve can lead to facial hemiparesis or paralysis, damage to vestibular structures to dizziness, and damage to the ear canal wall to cholesteatoma or meatal stenosis[16,17]. Consequently, operating surgeons need precise anatomical understanding and excellent motor skills. Mastoidectomy was commonly performed during the pre-antibiotic era to treat infections; sometimes >1,000 mastoidectomies per surgeon annually[18]. With antibiotics, mastoidectomy is less frequently performed but remains a cornerstone in otologic surgery, especially in the era of advanced surgical implants such as CIs[17].

The cochlea is a coiled, ~3–4 cm long bony tube filled with fluid, situated in the temporal bone[19]. It coils with ~2.75 turns, while the spiral ganglion inside only coils ~1.75 turns. Within the cochlea, three fluid-filled compartments are found: the scala vestibuli, media, and tympani. Delicate membranes separate these compartments: Reissner's/vestibular membrane is located between the scala vestibuli and media, and the basilar membrane between scala media and tympani. In cochlea's central axis (the modiolus), the spiral ganglion transmits nerve impulses. Normally, these nerve impulses originate from acoustic stimulation that creates a traveling wave, which stimulates the hair cells, and lead to an impulse in the cochlear nerve. The cochlea's central area is the preferred site of electrical stimulation. Consequently, optimal placement of the CI is inside the scala tympani near the modiolus[20]. The spiral ganglion is tonotopically organized: at cochlea's base, the highest frequencies are represented; frequencies lower progressively towards the apex[21]. Frequency range therefore depends on insertion depth and CI length.

A CI is an implanted hearing device converting sounds to electric signals, which are transferred via an electrode to the cochlea (Figure 1). It consists of 1) an external part comprising a microphone, sound processor, and transmitter, and 2) an internal part receiving the signal from the external part, forwarding it to an array of electrodes inside the cochlea. The first direct electrical stimulation of the human auditory system was in 1957[22]. An electrode was connected to the auditory nerve resulting in a rudimentary hearing function; yet, this implant failed. The next revolution was ~ten years later; nevertheless, CIs were simple, providing no speech perception. However, through the 1980s, this was attained via multipolar electrode stimulation at separate electrode contacts, each stimulating different frequencies. Today, CIs offer sufficient resolution for functional speech perception, allowing a near-normal lifestyle[23]. Eligibility criteria for CI vary by region and are expanding. Patients with profound sensorineural hearing loss with functional low-frequency hearing can benefit from CI in conjunction with acoustic stimulation (electric-acoustic stimulation)[24–26]. Careful CI placement is imperative and the *"electrode must be oriented appropriately (...) and a gentle insertion technique must be used (...)"*[27] Both apical stimulation, and CIs for perimodiolar placement (pre-curved, "modiolar hugging") may improve hearing at the cost of potentially increased intracochlear trauma[27,28]. Irrespective of electrode type and length, the foreign body (CI) in the scala tympani limits the traveling wave propagation of the basilar membrane, which—even without insertion trauma—leads to some degree of sensorineural hearing loss[29].

To achieve insertion, the cochlea is normally reached by a facial recess approach via a mastoidectomy and posterior tympanotomy.  Intracochlear access is gained by either 1) a cochleostomy where a hole is drilled in the cochlea; 2) an extended round window (RW) approach, where an anterior extension to the RW is drilled; or simply by 3) RW approach, where the bony overhang is removed, and a RW membrane slit made with a curved needle. The RW approach is preferable for minimizing the insertion trauma[30–32]. Regardless of approach, CI insertion entails some acute and/or long-term cochlear obstruction or damage[33,34]. Residual hearing is threatened when extensive trauma results from forceful or destructive insertion[35], underlining the significance of using slim and soft electrodes and a correct insertion technique[36,37].

**Figure 1:** Cochlear implant device. A: sound processor; B: external transmitter; C: implanted receiver; D: electrode array (© Oticon Medical)



## Acquiring surgical skills

### Apprenticeship learning (AL)

AL is the traditional way of learning surgery[38]. “See one, do one, teach one” is a distillation of its principle[39]: novices learn through direct supervision in the OR. Relying solely on AL is increasingly considered problematic: First, novices practice on actual patients. Stakeholders (e.g. patients, hospitals, and insurance companies) safety concerns question the concept of novice practice on patients[43,44]. Second, AL requires substantial costly supervision and OR time[40,41]. Third, AL occurs when patients undergo surgery, making time, location and procedure based on external circumstances rather than trainee needs[42]. Finally, residents used to work very long days but work hour restrictions have been implemented[45,46], making it harder to ensure sufficient surgical exposure[47]. Together, these changes have necessitated a re-evaluation of AL; simulation-based training (SBT) can remedy some of its problems[48].

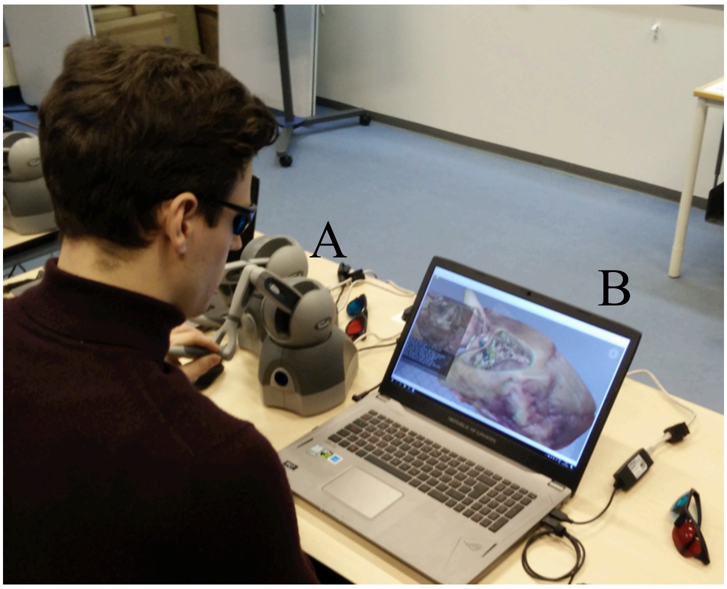
## Simulation-based training (SBT)

SBT aims to change *"see one, do one, teach one"* to *"see several, learn the skills during simulation training, do one, teach one"*[49], moving learning from patient to simulation. SBT is strongly supported by evidence compared with no training and mostly also compared with alternative instruction[50]. The effectiveness of mastoidectomy VR SBT was systematically reviewed, finding VR SBT skills to transfer to cadaver dissection[51]. Nonetheless, only two small studies evaluated effects of mastoidectomy SBT on real patient surgery, with no conclusive results[52,53].

### *Virtual reality temporal bone training—the Visible Ear Simulator*

Different VR simulators for otology exist[54]: CardinalSim (USA)[55], Ohio State University Simulator (USA)[56], University of Melbourne Simulator (Australia)[57], Visible Ear Simulator (VES; Denmark)[58], and VOXEL-MAN (Germany)[59]. In this thesis, we used the VES, developed by this thesis' principal supervisor, professor Mads Sølvsten Sørensen and computer scientist Peter Trier. The VES is academic freeware (i.e. free software for academic purposes) downloadable via the Internet, running on a PC with a Geforce GTX or RTX graphics card (Nvidia, Santa Clara, CA, USA). It features haptic force-feedback Geomagic Touch haptic device (3D Systems, Rock Hill, SC, USA)[60]. This setup (Figure 2) is relatively inexpensive and features a CI insertion module with both drilling and insertion (Figure 6).

**Figure 2:** VES setup. A: Geomagic Touch haptic device (3D Systems, Rock Hill, SC, USA); B: Gaming laptop running the simulation software



**Ericsson's deliberate practice**

Ericsson observed that experts intentionally address their shortcomings through *deliberate practice*: *"activities that have been specifically designed to improve the current level of performance"*[61]. It comprises nine elements: 1) Motivation; 2) Clear and relevant objective(s); 3) Suitable difficulty level; 4) Repeated practice; 5) High quality measurement of performance; 6) Pertinent feedback; 7) Error correction and monitoring; 8) Evaluation and comparison of performance with a set standard, and; 9) progression to the next practice unit[62]. Training should allow for deliberate practice by featuring these elements.

**A taxonomy of educational outcomes: Kirkpatrick's levels of hierarchy**

Educational outcomes are far from equally significant. For example, it is of no use to patients that trainees feel confident without improved treatment. Kirkpatrick's four-level appraisal framework addresses this[63]. Level 1 concerns reaction, e.g. learners' view of or satisfaction with training. Level 2a evaluates changes in attitudes/perception e.g. towards patients or others. Level 2b concerns acquisition of skills/knowledge. Level 3 describes changes in behavior, e.g. application/transfer of skills. Level 4a concerns organizational changes, e.g. changed care. Finally, level Level 4b covers improvements in health or well-being of patients[64].

## Learning curves

Learning curves (LC) relate effort (e.g. repetitions or time) with achieved learning[65]. Effort is represented by the x-axis; learning by the y-axis (Figure 5). Steep LC imply substantial training effect; conversely flat or plateauing LC mean no further improvement. Outcomes must be valid. For instance, "number of training procedures" (effort) and "objectively assessed surgical skills" (learning outcome), as used in Study II, are acceptable outcomes reflecting training amount and actual skills. Contrarily, surrogate measures such as "years of training" (effort) and "self-confidence" (skill) are not[66–68]. As learning occurs at varying paces, individual trainees' LC provide limited knowledge on general skills acquisition. Some learners quickly perform at a high level, whereas others' LC stall[69]. A normal LC shape is negatively accelerated, mirroring the fact that learners learn the most at the beginning of training[70]. Group LC can answer questions such as "when is learning most effective?" and "how much practice is needed?"[65]. Answering these questions is necessary for designing competence-based training.

## Transfer of skills

*Transfer of skills* is defined as *"application of knowledge and skills learned in one context to another"*[71], e.g. VR SBT skills to patients, 3D-printed models, or cadaver dissection. The higher the level of transfer in Kirkpatrick's hierarchy, the better;

ideally, to level 4B: improvements in patients' health or well being. Transfer is essential in SBT: acquiring skills, which only work during simulation without applicability in real-life is irrelevant. Further, SBT stakeholders need documentation that SBT works. SBT is based on assumption of transfer, especially from simulation to real-life; yet, few studies examine it with valid outcomes[72,73]. Relying on this assumption of transfer should not substitute systematic exploration to elucidate *how, why and when* transfer occurs and *how much* is transferred[73].

### Trainees' path to temporal bone surgery in Denmark

Subspecialization in otologic surgery occurs late in Denmark. The MD degree takes six years. One year of internship generally comprising medicine or surgery (never otorhinolaryngology (ORL)) and general practice/family medicine certifies for independent practice. Specialization can then slowly start by a one-year "introductory" specialty position. Next, 1–4 years of qualification for residency, e.g. training (including in related specialties) and/or research, is normally needed. ORL residency comprises four years of training without designated subspecialization. On average, specialization excluding MD degree takes 10.9 years[74,75]. Consistent exposure to temporal bone surgery normally commences after specialization at the earliest. This slow path to (sub)specialization explains the high age and low surgical experience in Study I, III, and IV's residents. For instance, the 18 residents in Study IV had a mean age of 34.5 years but were naïve to independently performing mastoidectomy.

## Instructional factors

Instructional factors are the "active ingredients or mechanism" in SBT[76]. Two systematic reviews identified effective instructional features[76,77]. Repetition, multiple learning strategies, and distributed practice (explained below) were effective. Three instructional features are described, which are pertinent for this thesis' studies.

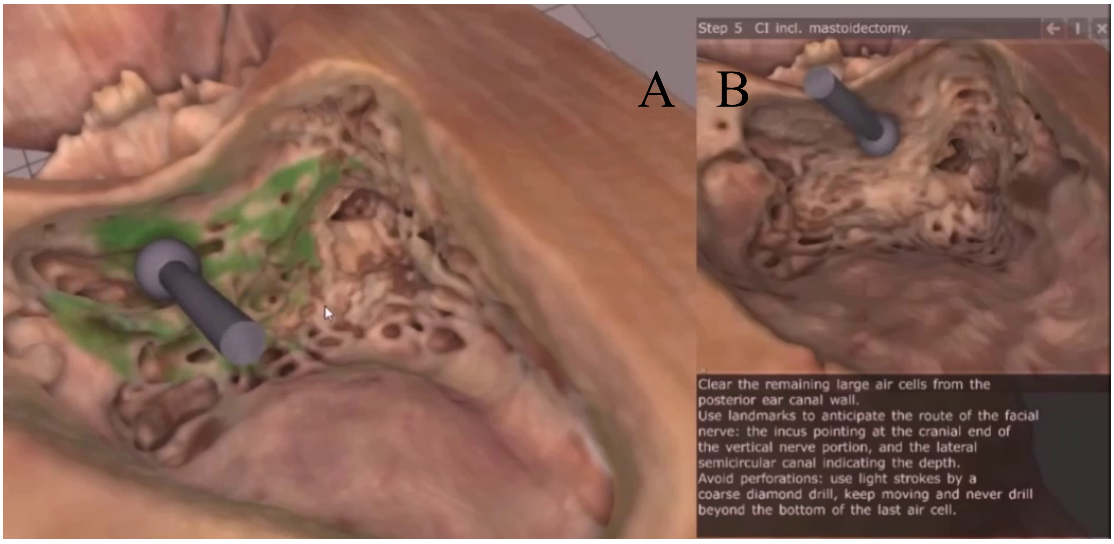
### Distribution of practice

Shorter training sessions separated by days (distributed practice) are superior to single, multi-hour sessions (massed practice)[78]. This is demonstrated in different procedures, including mastoidectomy[79,80]. The reason appears to be ample time for consolidation of skills, increasing cognitive capacity for the learning process[79,81,82].

### Directed, self-regulated learning (DSRL)

Training independently without direct supervision has been conceptualized using different terms[83–85]. In directed, self-regulated learning (DSRL), training is scaffolded by learning supports to guide learning via effective instructional design[86]. The VES features multiple learning supports for DSRL: on-screen dissection guides, simulator-integrated "green lighting" tutoring illustrating the volume to be drilled (Figure 3), alerts when injuring structures, structured self-assessment for formative feedback, and automated summative feedback[87–89].

**Figure 3:** Examples of learning supports in the VES. A: Simulator-integrated tutor function with green-lighting of the bone volume to be drilled during the dissection step[89]; B: On-screen dissection manual.



Clear the remaining large air cells from the posterior ear canal wall.
Use landmarks to anticipate the route of the facial nerve: the incus pointing at the cranial end of the vertical nerve portion, and the lateral semicircular canal indicating the depth.
Avoid perforations: use light strokes by a coarse diamond drill, keep moving and never drill beyond the bottom of the last air cell.

### Centralized and decentralized training

Technology-enhanced simulation is costly[90] due to high acquisition costs and need for introduction and support. As a result, it is often implemented in a simulation center (a centralized facility featuring equipment and staff[91–93]) and/or a “boot camp”, i.e. a course providing large training volume in a short time (massed practice)[94–96]. The popularity of “boot camp” training is paradoxical, since massed practice is inferior to distributed practice[97]. Decentralized training (DT) is a potential solution, also reported as *take-home training*[98], *home skills training*[99]*, off-site training*[86]*, portable simulation training*[100]*,* or *at-home training*[101]. DT can allow distributed training at home or in the local department. Disadvantages mirror those of DSRL: lack of hands-on instruction/feedback necessitating effective instructional design[102].

# Assessment of surgical skills

## Validity of assessment instruments

Surgical training should revolve around valid and reliable skills assessment. Key is the assessment instruments' validity, i.e. *"the extent to which the measurement method measures that, which it is intended to measure"*[103]. Assessment instrument must be supported by validity evidence for defendable use of scores[104].

Classical validity theory considers validity as four different types: face, content, criterion, and concurrent[105]. Some types seem arbitrary. For instance, "face validity" is a subjective measure of opinion, which provides limited validity[106], and can lead to wrong conclusions about physicians' competency[107]. Using face validity is therefore discouraged[108], although frequently used[55,109]. Further, classical validity theory does not evaluate test consequences.

Messick developed a validity framework featuring five sources of validity: 1) content (does test content reflect the construct of interest), 2) response process (bias elimination during response process/data collection), 3) internal structure (reliability), 4) relationship with other variables (relation between assessment scores and other variables), and 5) consequences (of the test)[110] (Table 1). Although this framework's validity appraisal is more systematic and thorough than classical validity theory[111] and supported by educational stakeholders[112], most validity studies use classical validity theory[109]. In this thesis, we used Messick's framework to gather validity evidence for the assessment tool developed in Study I.

### *Reliability: traditional reliability theory and generalizability theory*

Reliability—Messick's third source of validity—concerns consistency of test results. If the assessment is not reliable, limited inferences can be drawn from test scores[113].

Two approaches to reliability are used in Study I. Traditional reliability theory comprises calculation of internal consistency and inter-rater reliability. Internal consistency (Cronbach's alpha[114]) assesses whether test items reflect the same construct[115], while inter-rater reliability (e.g. intraclass correlation coefficient[116]) concerns consistency across raters. Generalizability theory estimates variance contribution from different variables (e.g. rater, participant, occasion etc.) quantified by a percentage of variance[117]. The generalizability coefficient is a measure of overall reliability;  >0.8 is considered sufficient for high-stakes assessment[118].

**Table 1** Validity evidence for the CISAT using Messick's framework

| **Source** | **Definition** | **Method** |
|---|---|---|
| Content | Relationship between the measured construct (CI-surgery) and content of test | Two clinical experts developed the content, targeting agreement between assessment tool content and CI surgery competence. |
| Response process | Consistency of response process for bias elimination | A single investigator (MF) was responsible for consistent data collection. Three blinded experts rated performances. |
| Internal structure | Consistency between test items and underlying construct | Cronbach's alpha assessed internal consistency of test scores; using generalizability theory, Generalizability- and Decision-studies evaluated reliability. |
| Relationship with other variables | Relationship between test scores and other variables (e.g. clinical competence) | Novices and CI-surgeons' performances were compared |
| Consequences | Impact and consequence of test scores | A pass/fail level was defined, allowing as few false positives (passed trained novices) and false negatives (failed CI-surgeons) as possible |

***Pass/fail standard setting***

Messick's fifth validity source is "consequences of testing" (Table 1). Tests often answer questions such as: "should the trainee progress to the next level?" or "is the trainee competent?”. Therefor, test scores must be operationalized with a pass/fail score. There are many different approaches to standard setting[119–124]. The contrasting groups’ method (Study I) compares groups of different skill levels[125] via performance Bell-curves. The curves’ intersection represents the pass/fail level with the fewest false positive and false negative.

### *Objective assessment of temporal bone surgery*

At least six different mastoidectomy assessment instruments exist; most lack validity evidence[126]. Final product assessment (FPA) uses final drilling results rather than constant rater presence throughout surgery[126]. The Welling Scale[127] is the FPA instrument supported by most validity evidence[126]; in Study IV, we used a modified version. Two assessment instruments for CI surgery exist: one for OR[128] and one for VR SBT[129]. Only the latter is relevant for VR SBT but unfortunately does not include electrode insertion; most content is included in assessment of mastoidectomy with posterior tympanotomy such as the Welling Scale[130]. As described, insertion aspects of CI surgery are pivotal for patient (hearing) outcomes, and drilling-only assessment seems simplistic. Further, outdated validity theory was used[129]. This lack of valid and reliable assessment of CI VR SBT is problematic because valid and reliable assessment is a prerequisite to evidence-based training.

## Research aims

*Study I*

To develop and gather validity evidence for an assessment tool for CI surgery using Messick's validity framework.

*Study II*

To determine the learning curve during CI VR SBT, and the transfer of cochlear implantation skills from VR SBT to a 3D-printed temporal bone.

*Study III*

To investigate the transfer of skills from CI VR SBT to cadaver CI surgery.

*Study IV*

To evaluate the feasibility and effect of a *decentralized training* VR SBT intervention on cadaver mastoidectomy performance.

# Research hypotheses

*Study I*

We hypothesized that validity evidence would support the use of the Cochlear Implant Surgery Assessment Tool (CISAT).

*Study II*

We hypothesized that learning curves described a typical, negatively accelerated pattern with a plateau of learning. Further, we hypothesized that skills obtained during VR SBT would transfer to improved performance on a 3D-printed temporal bone.

*Study III*

We hypothesized that the intervention group receiving CI VR SBT would outperform the control group, demonstrating transfer of skills from VR SBT to cadaver training.

*Study IV*

We hypothesized that decentralized mastoidectomy training would be feasible and could improve cadaver dissection performance.

# Summary of studies

## Overview

| | **Paper I** | **Paper II** | **Paper III** | **Paper IV** |
|---|---|---|---|---|
| Theme | Assessment of CI surgery skills | Learning curves of CI surgery | Transfer of skills in CI surgery | Decentralized virtual reality training |
| Design | Validity study using Messick's framework | Prospective interventional study | Randomized, controlled trial | Prospective, interventional cohort study |
| Number of participants | 35 | 24 | 18 | 36 |
| Participants | CI surgeons, ORL residents, medical students | Medical students | ORL residents | ORL residents |
| Raters | 3 | 2 | 2 | 2 |
| Assessment instrument | Cochlear Implant Surgery Assessment Tool (CISAT) | Modified Cochlear Implant Surgery Assessment Tool (mCISAT) | Cochlear Implant Surgery Assessment Tool (CISAT) | Modified Welling scale |
| Conclusion | The CISAT is supported by substantial validity evidence | CI VR SBT follows a traditional, negatively accelerated learning curve | CI VR SBT did not result in significant improvement of cadaver training performance | Decentralized VR SBT improves cadaver training performance |

Study III's participants comprise the control cohort in Study IV; Study III's intervention group comprises the ”trained novice” group in Study I

## Study I: Assessing competence in cochlear implant surgery using the newly developed Cochlear Implant Surgery Assessment Tool

### Background

CI surgery necessitates acquisition of excellent surgical skills. To measure these skills, an assessment instrument supported by validity evidence is needed. Such an assessment tool would be useful for reliable and valid assessment, e.g. during CI VR SBT. However, the existing assessment instruments for CI surgery are not suitable for CI VR SBT and are not supported by validity evidence gathered with modern methods[128,129]. We aimed to develop and collect validity evidence for a new assessment tool for CI surgery: the Cochlear Implant Surgery Assessment Tool (CISAT; Figure 2).

**Table 2** Cochlear Implant Surgery Assessment Tool (CISAT)

| | S | G | Performed by trainee with no or minimal guidance | | | | |
|---|---|---|---|---|---|---|---|
| 1. Posterior tympanotomy Posterior wall of auditory canal | □ | □ | 1<br>Substantial damage to vital structure (e.g. facial nerve), many remaining cells, holes in posterior wall. | 2 | 3<br>Few remaining cells, few holes in posterior wall, minor exposure of facial nerve | 4 | 5<br>Facial nerve just visible through layer of bone, no holes in posterior wall, no remaining cells |
| 2. Posterior tympanotomy: Initiating posterior tympanotomy | □ | □ | 1<br>Drilling at incorrect position, damaging facial nerve or tympanic cord | 2 | 3<br>Drilling near correct position and/or very minor nerve exposure | 4 | 5<br>Drilling close to incus buttress, ~2 mm fine diamond drill, no exposure of vital structures |
| 3. Posterior tympanotomy: Drilling technique | □ | □ | 1<br>Uncoordinated drill movement, inappropriate drill type/size | 2 | 3<br>Minor irrelevant movement and/or incorrect drill type/size | 4 | 5<br>Perfectly coordinated movement using correct drill type/size |
| 4. Posterior tympanotomy: Bone orientation for drilling | □ | □ | 1<br>No view of incus or lateral semicircular canal | 2 | 3<br>Partial overview and/or moderately incorrect viewing angle | 4 | 5<br>Incus buttress and lateral semicircular canal visible |
| 5. Posterior tympanotomy: Widening | □ | □ | 1<br>Insufficient widening or damage to adjacent structures | 2 | 3<br>Acceptable widening with minor nerve exposure | 4 | 5<br>Perfect widening laterally and anteriorly to facial nerve |
| 6. Posterior tympanotomy: Round window | □ | □ | 1<br>No round window membrane exposure | 2 | 3<br>Insufficient (<1 mm) or excessive round window membrane exposure | 4 | 5<br>Appropriate removal of round window bony overhang, ≥1 mm round window membrane exposure |
| 7.CI-insertion: Bone orientation at insertion | □ | □ | 1<br>No view of round window membrane | 2 | 3<br>Partial view of round window membrane | 4 | 5<br>Maximum view of round window membrane |
| 8. CI-insertion: Approach | □ | □ | 1<br>Electrode collisions, irregular movement, incorrect forceps grasp or path | 2 | 3<br>Few collisions, mostly fluid movement, partly correct grasp and path | 4 | 5<br>Deliberate movement with no collisions, perfect forceps grasp and path |
| 9. CI-insertion: Insertion vector | □ | □ | 1<br>Incorrect direction/vector of insertion | 2 | 3<br>Partly correct insertion vector | 4 | 5<br>Anterior, slightly lateral direction (avoiding hook region) |
| 10. CI-insertion: Speed and movement | □ | □ | 1<br>Insertion duration <5 s and/or abrupt, partially fast insertion | 2 | 3<br>5–15 s insertion duration and/or moderately smooth movement | 4 | 5<br>Continuous, smooth insertion, ≥15 s duration |
| 11. CI-insertion: Insertion result | □ | □ | 1<br>No insertion into cochlea | 2 | 3<br>Partial insertion (<3/4), electrode in scala tympani | 4 | 5<br>Complete scala tympani insertion |
| Overall assessment | | | Fail | | Borderline | | Pass |
| S: Performed by supervisor = 1 point; G: Performed with guidance = 2 points; s = seconds | | | | | | | |

## Methods

The Cochlear Implant Surgery Assessment Tool (CISAT; Table 2) was developed. Gathering of validity evidence followed Messick's validity framework, with five validity sources[131]. The first source, content, was addressed by having two experts in CI surgery define the substance of the assessment tool's items[132]. Two educational experts contributed to developing the assessment tool design. An early version of the CISAT was pilot tested and refined, leading to the final assessment tool. To gather validity evidence for the remaining sources of validity, three groups were enrolled: 1) novices (medical students), 2) trained novices (residents) who had completed two hours of CI VR SBT, and 3) CI surgeons, who had independently completed at least 25 CI procedures. Three blinded experts rated the performances. The consistency of the response process—Messick's second validity source—was ensured by having one investigator responsible for data collection, following a data collection protocol.

## Results

The final assessment tool comprised 11 items resulting in a crude total score of 11 to 55 points (i.e. including the 11 baseline points of to 1 point per 1–5-scored item). Reliability analysis—Messick's third validity source—revealed a generalizability coefficient of 0.76 and an excellent inter-rater reliability (intraclass correlation coefficient=0.92). The CISAT significantly discriminated between groups ($p<0.001$; Messick's fourth validity source).

To address Messick's fifth source of validity (consequences) two approaches to establishing a pass/fail level were explored. First, we used the contrasting groups' standard setting method[125], and set a pass/fail score of 36 points. As this pass/fail score allowed all trained novices to pass, a final pass/fail score of 45.3 points was established from the expert participants' mean score. Using this cutoff, no novices, 44% of trained novices, and 67% of experts would pass based on the observed performances in CI VR SBT.

**Conclusions**

The CISAT is reliable and supported by validity evidence. It can be used to measure the effect of training. The pass/fail score can be used to determine when to progress during training.

## Study II: Cochlear implant surgery: Learning curve in virtual reality simulation training and transfer of skills to a 3D-printed temporal bone—a prospective trial

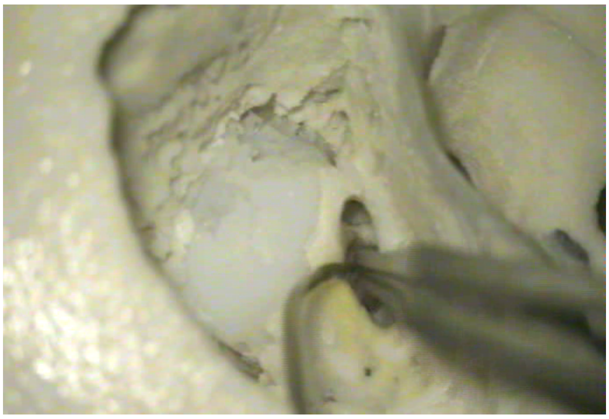
### Background

The learning curve describes the relationship between training effort and outcome. Outcomes should be evidence-based, i.e. based on competence rather than proxy measures such as time per procedure or self-evaluated skills[65]. Nonetheless, no credible studies have evaluated the learning curve of CI surgery. This gap in knowledge is problematic because knowing the learning curve is imperative when designing and planning meaningful training interventions. In this study, we aimed to evaluate the learning curve of CI VR training as well as the transfer of insertion skills to a 3D-printed temporal bone model.

### Methods

In this single-arm, prospective interventional trial, we recruited 24 novice medical students, who were naïve to temporal bone surgery and VR SBT. Before the CI VR SBT, participants completed a pre-test of insertion skills by performing a cochlear implantation on a pre-drilled 3D-printed temporal bone model (Model Schmidt, Phacon, Leipzig, Germany; Figure 4). Next, each participant performed 18 procedures during four training sessions (distributed practice). After the training intervention, participants completed a post-training test similar to the pre-test. Performances were scored by two blinded raters using a modified version of the CISAT (mCISAT).

**Figure 4:** Pre- and post training transfer test: CI insertion on a 3D-printed temporal bone model (Model Schmidt, Phacon, Leipzig, Germany).

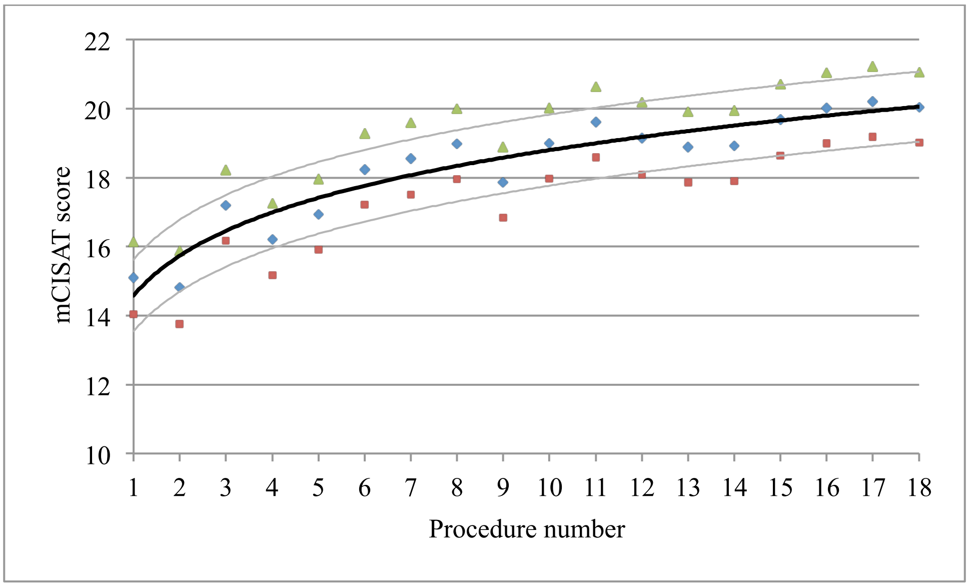


**Results**

The mean mCISAT score improved by 33% from 15.1 to 20.1 of a maximum of 28 points (p<0.001) during the training program. Evaluating the drilling and insertion related items separately, drilling improved by 43% (from 5.3 to 7.6 out of 12 points; p<0.001) whereas insertion improved by 28% (from 9.8 to 12.5 out of 16 points; p<0.001). Pre- and post training tests on a 3D-printed, pre-drilled temporal bone revealed a mean improvement of 21% (p<0.001). Learning curves were highly individual, but on average demonstrated a classical pattern with a decline of skills acquisition per procedure throughout the 18 procedures—i.e. a negatively accelerated learning curve. Nevertheless, a complete plateau of learning was not observed (Figure 5).

The rate of passing cochlear implantation performances on the 3D-printed temporal bone model (using the CISATs binary “Overall assessment”) were 44% at the pretest and 58% at the post test.

**Figure 5:** Mean scores during the 18 CI VR SBT procedures (black line). Grey lines represent 95% confidence intervals. Note that the y-axis includes only the range 10–22 of the modified CISAT's total range of 0–28 points.



## Conclusions

Reaching a consistent performance level—LC plateau—during CI VR SBT requires >18 repetitions, and continuous learning was observed even at the final repetition. Skills acquisition was highly individual. Skills improvement during transfer to a 3D-printed temporal bone model was modest; further studies should use models supported by validity evidence for training and assessment.

## Study III: Cochlear implant surgery: Virtual reality simulation training and transfer of skills to cadaver dissection—a randomized, controlled pilot study

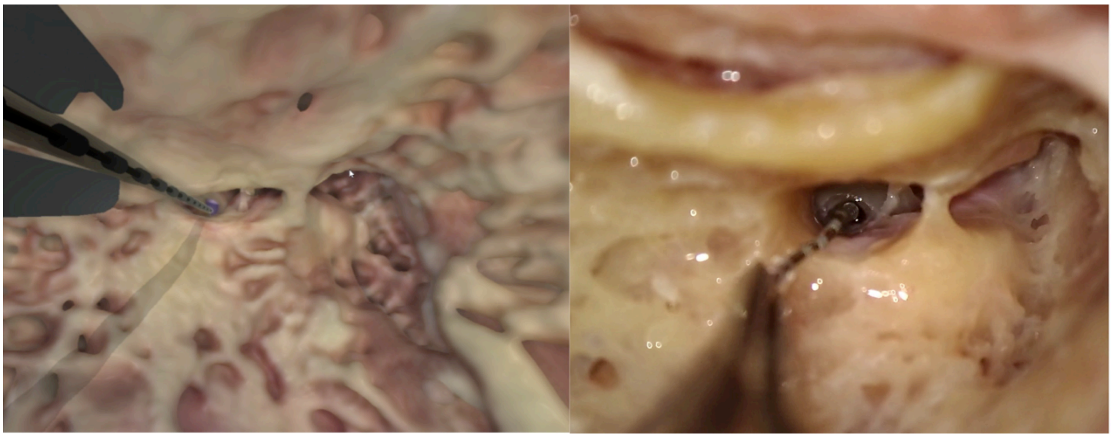
### Background

Virtual reality SBT of CI surgery is a new training option for acquiring basic skills in a risk-free environment with abundant training opportunity. Nonetheless, no study has evaluated the transfer of CI VR SBT to cadaver surgery and that is problematic since skills acquired during SBT are intended, but not guaranteed, to lead to improved OR performance. Knowing the transfer of skills from CI VR SBT to cadaver dissection would be useful to determine the value of CI VR SBT, in order for trainees to reach the highest possible level before cadaver training and patient surgery. Here we aimed to pilot the effect of CI VR SBT on subsequent cadaver dissection performance.

### Methods

This was a randomized, controlled pilot trial. Eighteen ORL residents were randomly assigned to the intervention comprising two hours of CI VR SBT (Figure 6) plus standard training (lectures and three hours of mastoidectomy VR SBT), or standard training only (controls). Drilling and insertion, as well the final drilling results were video recorded at the subsequent cadaver training (Figure 6). The videos were rated using the CISAT by two blinded experts. In addition, participants completed a structured questionnaire on the amount of assistance during cadaver dissection.

**Figure 6:** CI insertion (using the RW approach) in VR (left) and cadaver (right).



**Results**

The intervention group outperformed the control group by 5.4% (22.9 vs. 21.8 out of a maximum of 44 points when deducting the baseline score of 1 point per item) during cadaver training; this was not statistically significant (p=0.51). Evaluating the ability to perform the procedure independently, the intervention group received less assistance during the cadaver procedure (1.3 vs. 1.9 times; p=0.21). None of the main outcomes reached statistical significance.

**Conclusions**

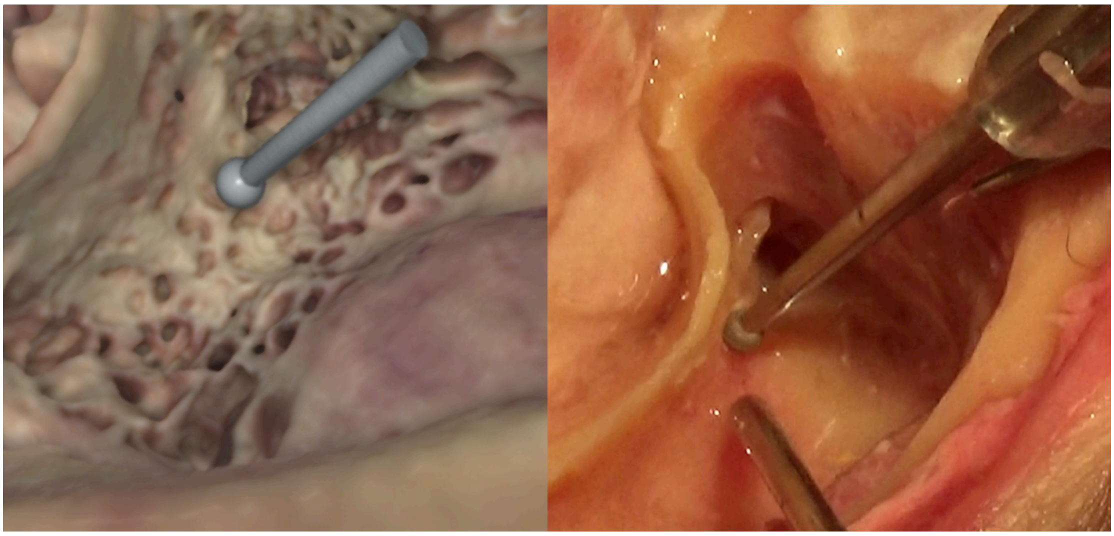
CI VR SBT is implementable in the context of a cadaver dissection course. Our findings indicate that CI VR SBT leads to a more self-directed cadaver surgery and also improved dissection performance slightly. The pilot study did not demonstrate a substantial training effect of the CI VR SBT intervention, suggesting either that the intervention had limited strength (for instance due to an insufficient training volume or simulator ineffectiveness), or that other methodological issues mask a potential effect. As such, the study adds to the literature by indicating that more CI VR SBT could be relevant before transition to cadaver dissection.

## Study IV: Decentralized Virtual Reality Training of Mastoidectomy Improves Cadaver Dissection Performance: A Prospective, Controlled Cohort Study

### Background

Virtual reality simulation training of temporal bone surgery is supported by substantial evidence; yet, there is a gap in knowledge on the best implementation in clinical practice[133]. In addition, most VR simulation training is conducted in a centralized facility (tertiary university hospital or simulation center), and often comprises massed practice, which is inefficient. Accordingly, there is a lack of knowledge that can inform the development of comprehensive curricula in temporal bone surgery. In this study, we aimed to evaluate the effect of a new instructional design for implementation of temporal bone VR training: as decentralized training in the trainees' local department or private home.

**Figure 7:** Drilling in VR (left) and subsequent cadaver dissection (right).



**Methods**

This was a prospective, controlled cohort study. Thirty-eight residents, who were generally novices in mastoidectomy, were enrolled: 20 in the intervention cohort and 18 in the control cohort. The intervention cohort was given the option to train decentralized supported by various learning supports for DSRL. They were not given any protected training time or other incentives to train. At subsequent cadaver dissection courses, both groups received standard training comprising lectures and three hours of VR simulation training of mastoidectomy. Finally, trainees performed a mastoidectomy (without posterior tympanotomy) on a human cadaver (Figure 7). Cadaver performances were rated by three blinded experts using a modified Welling Scale[134].

**Results**

Fifteen out of the 20 participants (75%) in the intervention cohort elected to train decentrally during the trial. On overage, participants in the intervention cohort trained 3.5 hours decentrally.

The intervention cohort scored a mean of 8.8 points during cadaver dissection, which was 76% more than the 5.0 point mean score in the control cohort ($p<0.001$).

**Conclusions**

Decentralized training is a new instructional design for implementing VR SBT of mastoidectomy. Our findings demonstrated that it was feasible to implement and use. Participants in the intervention cohort markedly outperformed participants in the control cohort during cadaver dissection, suggesting that the distributed, decentralized training led to substantial and clinically relevant skills improvement.

# Discussion

## Main findings

In this thesis on SBT, we first developed and collected validity evidence for a new assessment tool for CI surgery (CISAT; Study I). We found that validity evidence supports the CISAT, and that it is reliable in the context of CI VR SBT. Next, we evaluated acquisition of skills—the learning curve—during CI VR SBT and the transfer of CI insertion skills to a 3D-printed temporal bone model (Study II). Learning curves followed a negatively accelerated pattern with initial relatively substantial skills acquisition followed by a reduced return per repetition. The effect of VR training on subsequent insertion performance on a 3D-printed temporal bone was small. In Study III, we evaluated the effect of VR CI simulation training on cadaver dissection performance, finding that although the intervention cohort performed slightly better and needed less assistance during dissection, there was no significant group difference. Finally, in Study IV, we explored implementation of VR temporal bone SBT as *decentralized training* to gauge the feasibility and effect of training in a real-life setting without live instruction, supervision or protected training time. The intervention cohort performed substantially better than the control cohort during cadaver dissection.

## Skills assessment in CI VR surgery

The surgical assessment literature features many assessment instruments and their "validation"[109]. As described, most validity studies use outdated methods (classical validity theory) as opposed to contemporary validity frameworks such as Messick's[109]. Correspondingly, little is generally done to evaluate the consequence of the use of these assessment instruments—e.g. a pass/fail suggestion. This is the fifth validity source in Messick's framework, and important guidance for users of the assessment instruments' results[135,136]. Systematic reviews conclude that consequences are rarely explored in validity studies.[109,137]. Correspondingly, most studies provide no help for decision makers and other educational stakeholders on using and implementing assessments[50]. By suggesting a pass/fail score, we provided an estimate of the necessary level participants should consistently perform at before advancing during CI SBT. This could be used to determine when the trainee is qualified to progress to the next level of training, such as training on 3D-printed temporal bone models or cadavers.

Study I demonstrated significant differences between the three groups (novices, trained novices (residents), and CI surgeons), but there was a surprisingly small difference between trained novices and CI surgeons (42 vs. 45.3 CISAT points). This could suggests some degree of test construct underrepresentation[138,139] as the difference between the trained novices and real-life CI surgeons is orders of magnitude larger than these results suggest. It could be due to the assessment tool itself (poor discriminative ability beyond the novice level) or due to the VR simulation environment used (e.g. other properties than those applied during real-life CI surgery affect results achieved in the VR environment). For example, the ability to adapt to the VR environment could confound the results, making the novices appear better relative to the CI surgeons than they really are. In a validity study on the VOXEL-MAN temporal bone VR simulator, experts required significantly more time to get used to the simulator than residents, despite—or perhaps *because* of—their expertise in real-life temporal bone surgery[140]. In Study I, we gave all groups of participants a standardized warm-up. This is part of Messick's second validity source, *Response process,* i.e. minimizing bias during data collection. We did not offer experts more time to get used to the virtual environment and a longer warm-up period for these experts may have facilitated more representative performances from this group. Either way, it has been proposed that little weight should generally be put on expert-novice comparisons when evaluating assessment instruments' validity due to a high risk of confounding[141]. The only case in which exploring group differences makes a crucial difference in validity appraisal, is when no difference can be demonstrated between groups that are—judged by objective, external criteria—substantially different. As an example from temporal bone surgery research, Talk and colleagues attempted to "validate" an assessment tool for mastoidectomy (The Melbourne Mastoidectomy Scale) using classical validity theory. They found no significant difference between intermediates (residents) and experts[142]. This makes it hard to build a validity argument supporting the assessment scores using this tool. The modest difference between trained novices and experts could also suggest that the validity argument for the CISATs scores is strongest when evaluating novices. Although, as described, SBT is generally mostly targeted novices, this could be a problem for using the CISAT for CI VR SBT in trainees who are already competent performing the mastoidectomy procedure, which is the case for surgeons about to learn CI surgery. Nevertheless, if the CI-trainee's starting level is affected by the assessments' potentially poor discriminate ability beyond the novice level, it will have limited practical relevance because the CISAT might fail to detect progression from intermediate to expert. This potential problem is not unique to assessment of CI VR SBT: Jacobsen and colleagues attempted to address it by designing a test specifically for advanced cataract surgery; yet, despite the test being tailored for assessing non-novices, it did not succeed in reliably discerning intermediates from experts[143].

Study I has clear implications for clinical training: valid and reliable assessment of skills is a necessity if future training of CI surgery is to address the problems of AL described above by utilizing SBT for initial skills acquisition.
According to Ericsson, deliberate practice must be undertaken to attain expertise. As described, deliberate practice entails nine elements: 1) Motivation; 2) Goals; 3) Relevant difficulty level; 4) Repetition; 5) High quality performance measurement; 6) Feedback; 7) Error correction/monitoring; 8) Evaluation and comparison with a set level; 9) progression[62]. Study I potentially provides some of these elements: assessment of performance (element 5) can give the trainee evidence-based feedback on progression (element 6)—i.e. establish an individual learning curve useable for longitudinal evaluation of progress. Thereby, the trainee can set goals (element 2) for consistently reaching a set pass/fail standard (element 8). Also, working towards passing a certain skills level seems to increase motivation (element 1) during training[144]. In other words, Study I provides basis elements for deliberate practice and for gathering knowledge on skills progression in CI surgery.

In Study I, we encountered challenges when applying the contrasting groups' method, as the resulting cutoff-score of 36 out of 55 CISAT points would allow for all trained novices to pass. Such a low cutoff score is undesirable because it lets novices progress long before having exhausted the training potential of CI VR SBT. As a result, we instead utilized a different standard setting approach. Some suggest that simulation-based cutoff scores should exceed the intended clinical performance level because 1) simulation-based tests practically always involve some degree of construct underrepresentation, and 2) the stress and distractions of the clinical environment hamper performance compared with the safe, simulated environment[122,145]. The opposite can be said for the expert-group: here, the simulated environment is foreign and the experts underperform during simulation (compared with their true CI surgery abilities) because of dissimilarity with their usual real-life surgical environment[146]. We used the experts' mean as suggestion for a pass/fail score before progression e.g. to cadaver training[145]; an approach, which has previously been used in somewhat similar circumstances[147]. Using this pass/fail score would result in no novices, 44% of the trained novices and 67% of CI surgeons passing the test. The fact that we used two different approaches to standard setting after initially counterproductive result, underlines one of the fundamental aspects of standard setting: *"cut scores embody value judgments as well as technical and empirical considerations"*[124].

A core aspect of validity is the degree to which the content of the test reflects the actual construct of interest, in this case CI surgery; content is Messick's first validity source (Table 1). Experts frequently have differing opinions on the best procedural methods[148,149]. This is a problem when designing assessment instruments intended to work across different institutions and countries. To address this problem, consensus methods such as the Delphi methodology can sample a broad spectrum of opinions on assessment tool content[150,151]. A disadvantage of this methodology is that it can limit debate and discussion[152]. Also, conducting a Delphi survey can be time-consuming. In Study I, we instead elected to use a small group of four experts (two clinical and two educational), which allows for organic discussions and debate within a group acquainted with the setting where the assessment tool is to be used (CI VR SBT). However, as the CISAT is intended for use by an international audience, considering the opinions of two clinical experts to be ubiquitously relevant facts, and basing the assessment on it, could be problematic. Humphrey-Murto and colleagues argue that a main advantage of consensus methods is that it *"avoids undue dominance by specific individuals"*[152], and some might find that using a small group of experts from a single institution to determine assessment tool content results in just that. Balkany and colleagues developed a comprehensive, 38-item assessment tool for cadaver-based CI surgery. The authors gathered the content for the assessment tool by asking experts from different institutions for inputs. Unfortunately, they did not employ a modern validity framework but evaluated face validity. Further, they did not consider consequences of testing nor advanced reliability assessment[128]. This assessment tool's methods for gathering content validity is superior to the content validity of the CISAT; the opposite can be said for Messick's four remaining validity sources.

## Mapping skills acquisition in CI surgery

Monitoring or knowing the progression of learning a surgical procedure can aid both the individual trainee and educators. A knowledge gap existed in CI surgery, where no credible studies have explored the learning curve during training. Hence, the amount of training needed for proficiency is unknown. In Study II, we evaluated the learning curve for CI VR SBT and found that learning varied greatly among participants. It has been demonstrated that while some learners quickly perform competently, others follow a completely different and much flatter learning curve; others again seem unable to acquire the needed skills[153]. A wide range of levels was also observed in Study II, where learning curves among individual participants varied greatly. For example, mCISAT score (range 0–28, i.e. deducted the baseline score of 1 point per item) development from the first to the final procedure in the study included progressions such as 23→18.5 mCISAT points (-20%; -4.5 mCISAT points) to 6.5→24.5 (+276%; 18 mCISAT points).  This finding has implications for training, as it underlines the fact that objective assessment should be used instead of surrogate outcomes such as number of procedures or training time because these proxy measures do not correlate well with competency[65].

The finding that the passing rate (using the CISATs "Overall assessment") on the 3D-printed temporal bone model was 44% at the pretest and 58% at the post test seems surprising: a lower pre-test passing rate would be expected. This finding could indicate that new modalities, such as 3D-printed temporal bones for training and particularly testing should be used cautiously: the commercially available 3D-printed temporal bone model used (Model Schmidt, Phacon, Leipzig, Germany) is not supported by any meaningful validity evidence[154]. It appeared to be easier to insert the CI on the 3D-printed model than in any other modality (VR, cadaver, or live human), where the artificial "scala tympani", unrealistically, was wide and smooth. With no validity evidence supporting the model's use, it is hard to make definitive conclusions about transfer between VR SBT and 3D-printed models, or any other modality, in CI surgery based on Study II. In a recent systematic review, Frithioff and colleagues found that there is hardly any evidence in support of using 3D-printed temporal bone models[155]. Gathering such evidence is needed before further use of 3D-printed temporal bone models for training and assessment.

We included a substantial warm-up in Study II: two full procedures with extra learning supports. We considered this advantageous for assuring sufficient familiarization with the simulator. However, by including >1 hour of training as a warm-up, we likely induced learning before the actual start of the trial. Correspondingly, the baseline mean mCISAT score was relatively high at 15.2 out of

28 points (i.e. >50% of the maximum score at baseline). Three of the four participants scoring an even higher baseline mCISAT score >20 points improved <5%, demonstrating that they had already reached a relatively high performance level at the baseline measurement (>70% of the maximum score) and achieved no further improvement. This possible learning at baseline might reduce the observed impact of training.

We developed the CISAT (Study I) simultaneously to starting data collection for Study II. It was expected that the drilling would be appraised only by evaluating the final drilling result "final product" scoring as is the case e.g. with the Welling Scale. Nevertheless, once the CISAT and data collection for Study II were concluded, the CISAT also required video recordings of the drilling in addition to the final product assessment (CISAT items 2–4, Table 2). As a result, these CISAT items could not be evaluated in Study II, and the range of this modified CISAT (mCISAT) was reduced from 44 to 28 points (excluding the baseline score of one point per item). This has several implications: First, it means that the content of the assessment—Messick's first validity source—does not fully represent the construct of interest as identified by the experts who defined the content. Second, the reduced range means that the pass/fail score determined in Study I cannot be directly applied in Study II to estimate how much training is needed to pass the test (Messick's fifth validity domain). Third, the reduced range leads to a lower resolution of performance data.

No plateau of learning was observed when evaluating the overall learning curve in Study II: the participants still improved even at the final procedures. This suggests that more training is generally needed to reach a consistent performance level.

Mastery learning is an educational concept, which pivots on the achievement of *mastery* as guidance during training[76,156], rather than surrogate measures of competency such as number of procedures. In a mastery learning context, trainees must consistently perform at a set level before progressing to the next part of training or clinical practice. In Study I, we suggested such a standard (i.e. the pass/fail score), but did not use it in Study II, where participants trained a set number of repetitions rather than attaining consistently passing results. Andersen and colleagues evaluated the learning curve in mastoidectomy VR SBT using the VES, finding that participants undertaking distributed practice reached a plateau of learning after approximately nine repetitions[157]. Another study using the VOXEL-MAN simulator with integrated performance assessment identified this plateau even earlier at 4–5 repetitions[158]. These findings contrast those of Study II, where no clear plateau was observed. A clinical study on the learning curve of two doctors measuring surgery time and complications through a total of 98 procedures seemed to conclude that a stabile performance level was reached after 30 procedures[159]. It generally seems that CI-surgery is more complex to master than performing the preceding basic mastoidectomy, and therefore requires more practice.

## Moving initial learning from patients to simulation

For SBT to be relevant, skills obtained during SBT must transfer beyond the simulated setting. In Kirkpatrick's framework this corresponds to a higher outcome level. Kirkpatrick's level 2b concerns objectively measurable acquisition of skills[63,64]. No study had previously examined whether transfer occurs for CI VR SBT. In Study III, we sought to address this knowledge gap by evaluating transfer from CI VR SBT to cadaver training. Cadaver training is not equivalent to real patient surgery (lack of bleeding, tissue changes due to cadaver preservation, dissection lab rather than OR setting etc.), but is the best available substitute[54]. Study III demonstrated a surprisingly small and largely insubstantial performance improvement during cadaver surgery from CI VR SBT. As such, the study does little to support the hypothesis that CI VR SBT can move the initial part of the learning curve from real patient surgery (or cadaver surgery, which generally is of limited availibility[133]) to SBT.

Among the possible reasons for the lack of difference between the groups, two are the most probable: 1) training volume and 2) simulator ineffectiveness.

1) Training volume. Study II indicated that a large training volume (beyond 18 procedures/~9 hours) was relevant during CI VR SBT as learning still occurred; yet, in Study III, we gave the intervention group far less training (two hours). They completed a median 6.5 CI VR SBT procedures (range 2–8). The training volume in Study II was not based on attaining a set proficiency level (mastery learning[160]) but on feasibility relating to the implementation during the dissection course where the study took place. In Study I, we suggested a pass/fail score of 34.3 out of 44 points (corresponding to a crude score of 45.3 points including 1 baseline point per item); participants in Study III reached a mean score in their best CI VR SBT procedure of 30.4 points, i.e. below this predefined level. Utilizing Study II's findings on 1) the long learning curve and 2) absence of consistent performance during CI VR SBT (i.e. giving the intervention group more training) in combination with concepts of mastery learning (i.e. letting all participants train until consistently attaining a score of at least 34.3 points rather than training a set amount of time) would have aided in elucidating the true transfer effect.

2) Simulator ineffectiveness. In the surgical skills literature, the fidelity of training models was previously considered the main effectiveness-defining feature[88,161]. There is ongoing debate about the relationship between simulator fidelity and learning

efficacy[71,161,162], but most would agree that a certain degree of fidelity (or functional task alignment[71]) is necessary for skills transfer. In Study III, this concerns both the visual and haptic realism of the VES' CI module. Based on our findings, the VES version used might not reflect the CI procedure sufficiently accurately for measurable training outcomes when transferred to cadavers (Study III) or—for CI insertion—the 3D-printed temporal bone model used in Study II. In Messick's framework, this aspect is encompassed in the first validity domain: content. The content of the simulator or test should reflect the construct of interest (i.e. skills needed for real-life CI surgery). Study III's seemingly poor effect of CI VR SBT on cadaver dissection performance might suggest that there is misalignment between the skills learned during CI VR SBT and those needed during cadaver CI SBT[139]. Nevertheless, this is slightly contradicted by the findings of Study I where group comparisons revealed that CI surgeons markedly outperformed novices. Comparisons of groups such as novices and experts are popular in validity studies, and frequently mentioned as definitive evidence of validity. As described above, it is argued that they provide limited validity evidence and cannot not play a leading role in appraising validity[141].

## Implementing SBT in cochlear implant surgical training

Research on surgical SBT is generally focused on describing and evaluating new simulation methods[163] or, less frequently, theoretical considerations intending to expand or explain educational science[164]. However, implementing SBT into clinically feasible and meaningful training interventions is a different goal altogether that probably strikes a middle ground between theory and practice[165]. In a systematic review on self-guided learning, Brydges and colleagues concluded that future research should focus on *"(...) understanding of trainees' natural propensities while learning in the unsupervised context (...)"*. Using the concepts of directed, self-regulated learning (DSRL), we evaluated the feasibility and effect of implementing SBT using a new instructional design: as decentralized VR training (Study IV). The findings support the concept of DT in temporal bone surgery, as 1) DT was feasible to implement, 2) the intervention cohort outperformed the control cohort both statistically and clinically significantly. As such, the study adds to the current understanding of SBT in otology by demonstrating feasibility of this implementation in a clinical, everyday setting without protected training time. Nonetheless, we offered the intervention cohort training A+B (DT plus standard training) and the control cohort only A (standard training). Finding A+B to lead to better performance says nothing about how the DT compares with other types of training such as conventional centralized training[166]. Specifically for Study IV, the intervention cohort was given more mastoidectomy SBT, which is documented to improve cadaver dissection[80]. As such, Study IV does not compare effectiveness but simply suggests that DT is feasible and effective. Some argue that *whether* interventions work is less important than *why* they work, and that A vs. A+B studies are—from a medical education research standpoint—uninteresting[166,167].

In laparoscopic surgery, Fjørtoft and colleagues conducted a questionnaire study comprising 738 respondents within (general) surgery, gynecology, and urology, concluding that there is a gap in the implementation of SBT—an “implementation gap”—and proposed decentralized training as a possible solution[168]. Nevertheless, the literature on DT seems scattered, perhaps due to the terminological inconsistencies outlined above. Thinggaard and colleagues described a “take-home training” curriculum for basic laparoscopy skills using a “box trainer” i.e. a basic training platform for fundamental skills[169] . They found that residents individualized their training when decentralized training was available and that testing motivated them to train. A randomized, controlled trial comparing home training to simulation center training found home training to lead to more practice and better distribution of practice[170]. A trend towards better skills retention was also found. Another randomized trial on laparoscopy DT found pre-training hands-on introduction to be the main predictor for resident satisfaction with DT; residents randomized to training without it did not find the DT useful. This underlines the necessity of compatible instructional design during DT. Overall, residents enrolled in these studies had more immediate use for the basic laparoscopy skills during clinical practice than the residents in Study IV have for mastoidectomy. Additionally, Study IV utilized a VR rather than "box trainers". Consequently, the studies' results cannot be directly compared, but altogether suggest that 1) DT can allow for individualized training with better distribution of practice, 2) DT requires instructional design relevant for independent training, and 3) DT can yield better or equivalent results compared with traditional simulation center training.

The Welling Scale assessment tool used in Study IV differs from the CISAT/mCISAT used in Study I–III. One key difference concerns the response process (Messick's second validity source): the Welling Scale uses a checklist design, where completion of each part of the procedure is rated with a dichotomous score of one or zero. In contrast, the CISAT uses a rating scale design, where each item is given a score from 1–5, resulting in an increased resolution of scores. This potentially allows for a more nuanced assessment, and often better discriminative ability of skill levels, but increases subjectivity[171,172]. Despite the risk of increased subjectivity using rating scales, a meta-analysis found inter-rater reliability (encompassed in Messick's third validity source) largely similar[135]. Exploring the question of subjectivity in relation to reliability in this thesis' studies, the interrater reliability using the CISAT during cadaver dissection was 0.76 (intraclass correlation coefficient (ICC); two-way random for consistency[173]) in Study III (CISAT) and 0.88 in Study IV (modified Welling Scale). This corresponds to a moderate (CISAT) and good (modified Welling Scale) interrater reliability, respectively[173]. In contrast to the moderate reliability during cadaver dissection, the CISAT's interrater reliability during CI VR SBT was substantially higher (ICC=0.92; Study I). Finally, evaluating insertion in a 3D-printed temporal bone (Study II) resulted in a moderate interrater reliability (ICC=0.64). Altogether, these results mirror the finding that reliability of assessment in mastoidectomy is highly context dependent[174]. This has implications for training: if high stakes assessment using the CISAT is needed for cadaver (or live patient) surgery, more procedures and raters are needed than demonstrated in the generalizability theory analysis in Study I.

## Implications for clinical temporal bone training

The findings of this thesis can be used clinically to develop competence-based training in temporal bone surgery, i.e. to leverage the transition beyond the "see one, do one, teach on" approach of the apprenticeship model. The CISAT allows tracking of trainees' progression, especially for novices in temporal bone surgery. Training of CI surgery should be integrated into a curriculum for which this thesis' studies contribute several key aspects. With limited availability of cadaver specimens for gold-standard temporal bone SBT[133], the pass/fail standard setting can be used to evaluate when trainees have learned enough in CI VR SBT to use cadaveric temporal bones for further progression. A "VR SBT temporal bone certification" denoting consistently high performance scores might be used. This approach has been successfully used in ophthalmology[175].

Study II can be used when moving from the “one size fits all” approach in which the number of procedures or time of training is used as a surrogate for individual assessment of surgical competency: the findings demonstrate how the skills acquisition varies substantially between trainees. Further, the finding that a high practice volume is required can direct training of CI surgery: training should continue well beyond the <10 procedures where learning stagnates during mastoidectomy VR SBT.

Study III provides limited support for the transfer of skills from VR SBT to cadaver surgery, and for increased independence during cadaver dissection after CI VR SBT. The study’s implications for training mirror those of Study II, which found that learning continued beyond 18 procedures: the small training volume used in Study III (two hours) might be insufficient to yield significant results during transfer. A *mastery learning* training program, where all trainees continued VR SBT until reaching a predefined level before being allowed training on cadavers would address this problem.

Study IV has implications for training, as it illustrates an alternative to massed practice “boot camp” training, and demonstrates how decentralized training can also improve performance. In a period where centralized courses are hindered by COVID-19, this type of implementation seems more relevant than ever[176].

## Research perspectives

As described, we solely evaluated the CISAT in a simulated setting (VR, cadaver, and—for the insertion aspects—a 3D-printed temporal bone) and included only novices, trained novices (temporal bone surgery naïve residents), and CI surgeons. We did not evaluate how the CISAT performs in intermediates such as residents or fellows undertaking CI-surgery who are already proficient performing the basic mastoidectomy that precedes the surgery assessed by the CISAT. This "responsiveness" of the assessment tool i.e. *"the ability of an instrument to measure a meaningful or clinically important change in a clinical state"*[177] is essential for the clinical use of the CISAT within its target users and this knowledge gap should be addressed. In addition, there is a gap in knowledge about the effect of CI SBT, i.e. "sensitivity to change" in this group of learners[178]. Knowing *when*, *how*, and for *whom* CI VR SBT and the CISAT works—especially concerning intermediates—would be helpful to implement evidence-based CI SBT. The research questions to answer are: 1) *Can validity evidence support using the CISAT in aspiring CI surgeons who are proficient in performing a basic mastoidectomy?* and 2) *What is the effect of SBT in this group?*

SBT is not only relevant for acquiring skills; it can also play a role in skills retention in surgeons who do not perform the procedure frequently[179]. There is a gap in knowledge, as no study to date evaluated retention of skills in CI surgery. Such knowledge would be useful: As more regions offer CI, patient eligibility criteria continue to expand, and demographic changes increase the demand for CIs, more surgeons will need to learn and perform CI surgery. The initial volume per surgeon might be insufficient for natural retention of skills. SBT could aid skills retention, and this would be relevant to investigate, e.g. by using CI VR SBT or 3D-printed models[155]. The research question that needs answering is: *Can CI SBT improve skills retention in CI surgery?*

Finally, a lack of knowledge about the transfer of skills from CI VR SBT remains. Due to reasons described above, we did not succeed in filling this gap with Study III. An improved study featuring 1) mastery learning principles to ensure that all participants train CI VR SBT to proficiency or beyond[145], rather than a set number of SBT procedures 2) a refined iteration of the VES' CI module, and 3) a larger number of participants, is needed to unravel the potential of CI VR SBT transfer.

## Limitations: external validity and design

### *External validity*

Surgeons who are in a process of learning temporal bone surgery are likely more motivated and have a vastly stronger incentive to learn than medical students or residents without immediate use for the skills acquired in the studies. In accordance with Ericsson's observations, motivation is pivotal for acquiring skills[180] and motivation is a strong predictor for skills acquisition in mastoidectomy training[181]. The likely modest motivation—or at least unfavorable incentive structure—in the study participants (medical students and residents not bound for imminent independent temporal bone surgery) makes it challenging to make definitive informed decisions about the target learners' potential effect of a similar training intervention, and could lead to underestimation of the training effect. On the other hand, the influence of being observed and tested (the Hawthorne effect) might skew our observations in the opposite direction[182].

### *Design*

It is hard to conceive a perfect study design that is also feasible; rather, study designs are often a compromise between methodological aspirations and feasibility. Nonetheless, there are limitations concerning the design of the studies in this thesis. Study II utilized a single-group pretest-posttest design to evaluate the transfer of skills to a 3D-printed model. Cook argues that *"(...)this design is susceptible to numerous validity threats including history, maturation, testing, instrumentation, regression, location, and attitude. Collectively, these threats seriously constrain the inferences that can be drawn from research using this design."*[183] The pre-test increases statistical power, but also primes participants for the post training transfer test (i.e. learning by testing), thereby reducing the generalizability of the findings relating to the transfer test.

Study III was statistically powered to detect only a large difference between groups (>30%); more statistical power (i.e. more participants) would be favorable. Further, although randomized controlled trials are considered the highest level of evidence[184], comparing something to nothing—which we did in both Study III and IV—is regarded irrelevant by many medical education researchers[183,185] because training practically always leads to learning. Ironically, Study III did not demonstrate such learning. In attempting to gain statistical power lacking in Study III, we utilized a prospective cohort design in Study IV. This adds statistical power, but introduces a new problem: cohort differences. There were no clear differences in baseline characteristics (e.g. demographics) between the cohorts, but other, seemingly intangible group differences might still apply as some groups of learners are simply more proactive than others. This phenomenon of group dynamics' sometimes profound effects on learning has been demonstrated extensively[186] but can be challenging to account for in cohort studies such as Study IV.

## Conclusion

We developed an assessment instrument for CI SBT and explored skills assessment, learning curves, and transfer during CI VR SBT. We found that the CISAT was reliable and supported by validity evidence from Messick's five validity sources. Learning curves followed a negatively accelerated pattern without reaching a plateau, suggesting that CI VR SBT has a longer learning curve before plateauing than mastoidectomy. In a randomized, controlled trial, two hours of CI VR SBT did not lead to significantly improved performance during cadaver training. Finally, we tested the feasibility and effect of a new instructional design implementation of mastoidectomy VR SBT—decentralized training—that was viable and effective.

The thesis bridges several knowledge gaps in temporal bone VR SBT—knowledge that is directly applicable in training curricula. Altogether, continued innovations in CI, increasing worldwide demand with widening patient eligibility criteria, and steadily rising awareness of ensuring surgeons' competencies before operating patients suggest that there is a clinical potential for evidence-based training of CI surgery. This thesis represents a small step on the path towards such evidence-based training.

# References

1. Goman, A. M. & Lin, F. R. Prevalence of hearing loss by severity in the United States. *Am. J. Public Health* **106**, 1820–1822 (2016).
2. Turton, L. & Smith, P. Prevalence & characteristics of severe and profound hearing loss in adults in a UK National Health Service clinic. *Int. J. Audiol.* **52**, 92–97 (2013).
3. Blanchfield, B. B., Feldman, J. J. & Dunbar, J. The severely to profoundly hearing impaired population in the United States: prevalence and demographics. *Policy Anal. Brief. H Ser.* **1**, 1–4 (1999).
4. Stevens, G. *et al.* Global and regional hearing impairment prevalence: an analysis of 42 studies in 29 countries. *Eur. J. Public Health* **23**, 146–152 (2013).
5. World Health Organisation. Deafness and hearing loss. (2020). Available at: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss. (Accessed: 3rd February 2021)
6. Bichey, B. G. & Miyamoto, R. T. Outcomes in bilateral cochlear implantation. *Otolaryngol. - Head Neck Surg.* **138**, 655–661 (2008).
7. Svinndal, E. V., Solheim, J., Rise, M. B. & Jensen, C. Hearing loss and work participation: a cross-sectional study in Norway. *Int. J. Audiol.* **57**, 646–656 (2018).
8. Hogan, A., O'Loughlin, K., Davis, A. & Kendig, H. Hearing loss and paid employment: Australian population survey findings. *Int. J. Audiol.* **48**, 117–122 (2009).
9. Impairments, N. R. C. (US) C. on D. D. for I. with H., Dobie, R. A. & Hemel, S. Van. Impact of Hearing Loss on Daily Life and the Workplace. (2004).

10. Mohr, P. E. *et al.* The societal costs of severe to profound hearing loss in the United States. *Int. J. Technol. Assess. Health Care* **16**, 1120–1135 (2000).

11. World Health Organisation. *Global costs of unaddressed hearing loss and cost-effectiveness of interventions*. (2017).

12. Eshraghi, A. A. *et al.* The cochlear implant : Historical aspects and future prospects. **295**, 1967–1980 (2016).

13. Brennan, T. A. *et al.* Incidence of Adverse Events and Negligence in Hospitalized Patients. *N. Engl. J. Med.* **324**, 370–376 (1991).

14. Jeppesen, J. & Faber, C. E. Surgical complications following cochlear implantation in adults based on a proposed reporting consensus. *Acta Oto-Laryngologica* **133**, 1012–1021 (2013).

15. Fitts, P. M. & Posner, M. I. *Human performance. Human performance.* (Brooks/Cole, 1967).

16. Tange, R. A., Grolman, W. & Maat, A. Intracochlear misdirected implantation of a cochlear implant. *Acta Otolaryngol.* **126**, 650–652 (2006).

17. Theissing, J., Rettinger, G. & Werner, J. A. *ENT—Head and Neck Surgery: Essential Procedures*. (Thieme, 2010).

18. Bento, R. F. & De Oliveira Fonseca, A. C. A brief history of mastoidectomy. *International Archives of Otorhinolaryngology* **17**, 168–178 (2013).

19. Rask-Andersen, H. *et al.* Human cochlea: Anatomical characteristics and their relevance for cochlear implantation. *Anat. Rec.* **295**, 1791–1811 (2012).

20. Khurayzi, T., Dhanasingh, A., Almuhawas, F. & Alsanosi, A. Shape of the Cochlear Basal Turn: An Indicator for an Optimal Electrode-to-Modiolus Proximity With Precurved Electrode Type. *Ear, Nose Throat J.* **100**, 38–43 (2021).

21. Shera, C. A. The spiral staircase: Tonotopic microstructure and cochlear tuning. *J. Neurosci.* **35**, 4683–4690 (2015).

22. Wilson, B. S. & Dorman, M. F. Cochlear implants: A remarkable past and a brilliant future. *Hear. Res.* **242**, 3–21 (2008).

23. Holden, L. K. *et al.* Factors affecting open-set word recognition in adults with cochlear implants. *Ear Hear.* **34**, 342–360 (2013).

24. Moteki, H. *et al.* Long-term results of hearing preservation cochlear implant surgery in patients with residual low frequency hearing. *Acta Otolaryngol.* **137**, 516–521 (2017).

25. Usami, S.-I. *et al.* Hearing preservation and clinical outcome of 32 consecutive electric acoustic stimulation (EAS) surgeries. *Acta Otolaryngol.* **134**, 717–727 (2014).

26. Von Ilberg, C. A., Baumann, U., Kiefer, J., Tillein, J. & Adunka, O. F. Electric-acoustic stimulation of the auditory system: A review of the first decade. *Audiology and Neurotology* **16**, 1–30 (2011).

27. Luxford, W. L. & Cullen, R. D. *Surgery for Cochlear Implantation*. *Otologic Surgery* (Elsevier, 2010). doi:10.1016/B978-1-4160-4665-3.00031-7

28. Eshraghi, A. A., Yang, N. W. & Balkany, T. J. Comparative study of cochlear damage with three perimodiolar electrode designs. *Laryngoscope* **113**, 415–419 (2003).

29. Semmelbauer, S., Böhnke, F. & Müller, J. M. Influence of a cochlea implant electrode on the traveling wave propagation. in *AIP Conference Proceedings* **1965**, 050001 (American Institute of Physics Inc., 2018).

30. Adunka, O. *et al.* Cochlear implantation via the round window membrane minimizes trauma to cochlear structures: A histologically controlled insertion

study. *Acta Otolaryngol.* **124**, 807–812 (2004).

31. Jiam, N. T., Jiradejvong, P., Pearl, M. S. & Limb, C. J. The effect of round Windowvs cochleostomy surgical approaches on cochlear implant electrode position a flat-panel computed tomography study. *JAMA Otolaryngol. - Head Neck Surg.* **142**, 873–880 (2016).

32. Eshraghi, A. A. *et al.* Clinical, surgical, and electrical factors impacting residual hearing in cochlear implant surgery. *Acta Otolaryngol.* **137**, 384–388 (2017).

33. Eshraghi, A. A. *et al.* Mechanisms of programmed cell death signaling in hair cells and support cells post-electrode insertion trauma. *Acta Otolaryngol.* **135**, 328–334 (2015).

34. Richard, C., Fayad, J. N., Doherty, J. & Linthicum, F. H. Round window versus cochleostomy technique in cochlear implantation: Histologic findings. *Otol. Neurotol.* **33**, 1181–1187 (2012).

35. Gantz, B. J. & Turner, C. W. Combining acoustic and electrical hearing. *Laryngoscope* **113**, 1726–1730 (2003).

36. Finley, C. C. *et al.* Role of electrode placement as a contributor to variability in cochlear implant outcomes. *Otol. Neurotol.* **29**, 920–8 (2008).

37. Coordes, A., Ernst, A., Brademann, G. & Todt, I. Round window membrane insertion with perimodiolar cochlear implant electrodes. *Otol. Neurotol.* **34**, 1027–1032 (2013).

38. Hamdorf, J. M. & Hall, J. C. Acquiring surgical skills. *Br. J. Surg.* **87**, 28–37 (2000).

39. Akhtar, K. S. N., Chen, A., Standfield, N. J. & Gupte, C. M. The role of simulation in developing surgical skills. *Curr. Rev. Musculoskelet. Med.* **7**,

155–160 (2014).

40. Jabbour, N. & Snyderman, C. H. The Economics of Surgical Simulation. *Otolaryngologic Clinics of North America* **50**, 1029–1036 (2017).

41. Bridges, M. & Diamond, D. L. The financial impact of teaching surgical residents in the operating room. *Am. J. Surg.* **177**, 28–32 (1999).

42. Kyser, K. L. *et al.* Forceps delivery volumes in teaching and Nonteaching hospitals: Are volumes sufficient for physicians to acquire and maintain competence? *Acad. Med.* **89**, 71–76 (2014).

43. Harper, P. R. & Pitt, M. A. On the challenges of healthcare modelling and a proposed project life cycle for successful implementation On the challenges of healthcare modelling and a proposed project life cycle for successful implementation $. *J. Oper. Res. Soc.* **55**, 657–661 (2004).

44. Brailsford, S. C., Bolt, T., Connell, C., Klein, J. H. & Patel, B. Stakeholder engagement in health care simulation. in *Proceedings - Winter Simulation Conference* 1840–1849 (2009). doi:10.1109/WSC.2009.5429190

45. Antiel, R. M. *et al.* ACGME Duty-Hour Recommendations — A National Survey of Residency Program Directors. *N. Engl. J. Med.* **363**, e12 (2010).

46. Nasca, T. J., Day, S. H. & Amis, E. S. The New Recommendations on Duty Hours from the ACGME Task Force. *N. Engl. J. Med.* **363**, e3 (2010).

47. Chikwe, J., De Souza, A. C. & Pepper, J. R. No time to train the surgeons. *British Medical Journal* **328**, 418–419 (2004).

48. Kneebone, R. Evaluating clinical simulations for learning procedural skills: A theory-based approach. *Academic Medicine* **80**, 549–553 (2005).

49. Bashankaev, B., Baido, S. & Wexner, S. D. Review of available methods of simulation training to facilitate surgical education. *Surg. Endosc.* **25**, 28–35

(2011).

50. Cook, D. A. How much evidence does it take? A cumulative meta-analysis of outcomes of simulation-based education. *Med. Educ.* **48**, 750–760 (2014).

51. Lui, J. T. & Hoy, M. Y. Evaluating the Effect of Virtual Reality Temporal Bone Simulation on Mastoidectomy Performance: A Meta-analysis. *Otolaryngology - Head and Neck Surgery (United States)* **156**, 1018–1024 (2017).

52. Al-Noury, K. Virtual Reality Simulation in Ear Microsurgery: A Pilot Study. *Indian J. Otolaryngol. Head Neck Surg.* **64**, 162–166 (2012).

53. Gawęcki, W. *et al.* The Impact of Virtual Reality Training on the Quality of Real Antromastoidectomy Performance. *J. Clin. Med.* **9**, 3197 (2020).

54. Bhutta, M. F. A review of simulation platforms in surgery of the temporal bone. *Clin. Otolaryngol.* **41**, 539–545 (2016).

55. Compton, E. C. *et al.* Assessment of a virtual reality temporal bone surgical simulator: A national face and content validity study. *J. Otolaryngol. - Head Neck Surg.* **49**, (2020).

56. Wiet, G. J. *et al.* Virtual temporal bone dissection system: OSU virtual temporal bone system: Development and Testing. *Laryngoscope* **122**, 1–12 (2012).

57. O'Leary, S. J. *et al.* Validation of a Networked Virtual Reality Simulation of Temporal Bone Surgery. *Laryngoscope* **118**, 1040–1046 (2008).

58. Sorensen, M. S., 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.* **30**, 484–487 (2009).

59. Wilmot, V., Bennett, W., Reddy, V. & Alderson, D. Comparative study of the

realism and usefulness of two virtual temporal bone simulators: The voxel-man temposurg and the visible ear simulator. *Int. J. Surg.* **12**, S38 (2014).

60. Sørensen, M. S., Mikkelsen, P. T. & Andersen, S. A. W. Visible Ear Simulator download page. Available at: https://ves.alexandra.dk/forums/ves3-ready.

61. Ericsson, K. A. *et al.* The Role of Deliberate Practice in the Acquisition of Expert Performance. **100**, 363–406 (1993).

62. Patnaik, R. & Stefanidis, D. Outcome-Based Training and the Role of Simulation. in *Comprehensive Healthcare Simulation* 69–78 (Springer, Cham, 2019). doi:10.1007/978-3-319-98276-2_7

63. Kirkpatrick, D. L. & Kirkpatrick, J. D. *Evaluating Training Programs: The Four Levels*. **1**, (Berrett-Koehler, 1994).

64. Hammick, M. Interprofessional education: Evidence from the past to guide the future. *Med. Teach.* **22**, 461–467 (2000).

65. Pusic, M. V., Boutis, K., Hatala, R. & Cook, D. A. Learning Curves in Health Professions Education. *Acad. Med.* **90**, 1034–1042 (2015).

66. Eva, K. W. & Regehr, G. Self-Assessment in the Health Professions: A Reformulation and Research Agenda. *Acad. Med.* **80**, S46–S54 (2005).

67. Colliver, J. A., Verhulst, S. J. & Barrows, H. S. Self-assessment in medical practice: A further concern about the conventional research paradigm. *Teaching and Learning in Medicine* **17**, 200–201 (2005).

68. Eva, K. W. & Regehr, G. Exploring the divergence between self-assessment and self-monitoring. *Adv. Heal. Sci. Educ.* **16**, 311–329 (2011).

69. Pusic, M., Pecaric, M. & Boutis, K. How much practice is enough? Using learning curves to assess the deliberate practice of radiograph interpretation. *Acad. Med.* **86**, 731–736 (2011).

70. Thurstone, L. L. The learning curve equation. *Psychol. Monogr.* **26**, i–51 (1919).

71. Hamstra, S. J., Brydges, R., Hatala, R., Zendejas, B. & Cook, D. A. Reconsidering fidelity in simulation-based training. *Acad. Med.* **89**, 387–392 (2014).

72. Sturm, L. P. *et al.* A systematic review of skills transfer after surgical simulation training. *Ann. Surg.* **248**, 166–179 (2008).

73. Dawe, S. R. *et al.* Systematic review of skills transfer after surgical simulation-based training. *Br. J. Surg.* **101**, 1063–1076 (2014).

74. Det Nationale Institut for Kommuners og Regioners Analyse og Forskning. *Medicinstuderende og yngre laegers speciale-og karrierevalg*. (2009).

75. Sundhedsstyrelsen. *Speciallægeuddannelsen – en status og perspektivering*. (2012).

76. Cook, D. A. *et al. Comparative effectiveness of instructional design features in simulation-based education: Systematic review and meta-analysis*. *Medical Teacher* **35**, (2013).

77. Issenberg, S. B., McGaghie, W. C., Petrusa, E. R., Gordon, D. L. & Scalese, R. J. Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. *Med. Teach.* **27**, 10–28 (2005).

78. Moulton, C. A. E. *et al.* Teaching surgical skills: What kind of practice makes perfect? A randomized, controlled trial. *Ann. Surg.* **244**, 400–407 (2006).

79. Andersen, S. A. W., Mikkelsen, P. T., Konge, L., Cayé-Thomasen, P. & Sørensen, M. S. Cognitive load in distributed and massed practice in virtual reality mastoidectomy simulation. *Laryngoscope* **126**, E74–E79 (2016).

80. Andersen, S. A. W., Foghsgaard, S., Cayé-Thomasen, P. & Sørensen, M. S.

The Effect of a Distributed Virtual Reality Simulation Training Program on Dissection Mastoidectomy Performance. *Otol. Neurotol.* **39**, 1277–1284 (2018).

81. Brashers-Krug, T., Shadmehr, R. & Bizzi, E. Consolidation in human motor memory. *Nature* **382**, 252–255 (1996).

82. Shea, C. H., Lai, Q., Black, C. & Park, J. H. Spacing practice sessions across days benefits the learning of motor skills. *Hum. Mov. Sci.* **19**, 737–760 (2000).

83. Pilling-Cormick, J.; Garrison, R. Self-directed and self-regulated learning: Conceptual links. *Can. J. Univ. …* **33**, 13–33 (2007).

84. Brydges, R. *et al.* Self-regulated learning in simulation-based training: A systematic review and meta-analysis. *Med. Educ.* **49**, 368–378 (2015).

85. Boekaerts, M. *Self-regulated learning: where we are today. M. Boekaerts) International Journal of Educational Research* **31**, (1999).

86. Thinggaard, E. *et al.* Off-site training of laparoscopic skills, a scoping review using a thematic analysis. *Surg. Endosc.* **30**, 4733–4741 (2016).

87. Chen, J. *et al.* Use of Automated Performance Metrics to Measure Surgeon Performance during Robotic Vesicourethral Anastomosis and Methodical Development of a Training Tutorial. *J. Urol.* **200**, 895–902 (2018).

88. Cook, David A, Hatala, Rose, brydges, Ryan, Zendjas,  benjamin. Technology-Enhanced Simulation for Health Professions Education A Systematic Review and Meta-analysis. *JAMA J. Am. Med. Assoc.* **306**, 978–988 (2011).

89. Andersen, S. A. W., Mikkelsen, P. T. & Sørensen, M. S. The Effect of Simulator-Integrated Tutoring for Guidance in Virtual Reality Simulation Training. *Simul. Healthc. J. Soc. Simul. Healthc.* **15**, 147–153 (2020).

90. Kapadia, M. R., DaRosa, D. A., MacRae, H. M. & Dunnington, G. L. Current

Assessment and Future Directions of Surgical Skills Laboratories. *J. Surg. Educ.* **64**, 260–265 (2007).

91. MacRae, H. M., Satterthwaite, L. & Reznick, R. K. Setting up a surgical skills center. *World J. Surg.* **32**, 189–195 (2008).

92. Meier, A. H. Running a surgical education center: From small to large. *Surgical Clinics of North America* **90**, 491–504 (2010).

93. Konge, L. *et al.* The Simulation Centre at Rigshospitalet, Copenhagen, Denmark. *J. Surg. Educ.* **72**, 362–365 (2015).

94. Sonnadara, R., Mironova, P. & Rambani, R. *Boot Camp Approach to Surgical Training*. *Boot Camp Approach to Surgical Training* (2018). doi:10.1007/978-3-319-90518-1

95. Neylan, C. J. *et al.* Medical School Surgical Boot Camps: A Systematic Review. *J. Surg. Educ.* **74**, 384–389 (2017).

96. Malekzadeh, S., Deutsch, E. S. & Malloy, K. M. Simulation-based otorhinolaryngology emergencies boot camp: Part 2: Special skills using task trainers. *Laryngoscope* **124**, 1566–1569 (2014).

97. Weis, J. J., Farr, D., Abdelfattah, K. R., Hogg, D. & Scott, D. J. A proficiency-based surgical boot camp May not provide trainees with a durable foundation in fundamental surgical skills. *Am. J. Surg.* **217**, 244–249 (2019).

98. Wilson, E. *et al.* Improved laparoscopic skills in gynaecology trainees following a simulation-training program using take-home box trainers. *Aust. New Zeal. J. Obstet. Gynaecol.* 1–7 (2018). doi:10.1111/ajo.12802

99. Harvey, L. F. B., King, L. & Hur, H.-C. Challenges Associated with a Self-Contained Simulation Curriculum Using a Home Laparoscopic Skills Trainer. *J. Minim. Invasive Gynecol.* **20**, S130–S131 (2013).

100. Nakamura, L. Y. *et al.* Comparing the portable laparoscopic trainer with a standardized trainer in surgically naïve subjects. *J. Endourol.* **26**, 67–72 (2012).

101. Sharpe, B. A., Machaidze, Z. & Ogan, K. Randomized comparison of standard laparoscopic trainer to novel, at-home, low-cost, camera-less laparoscopic trainer. *Urology* **66**, 50–54 (2005).

102. de Jong, T. Cognitive load theory, educational research, and instructional design: Some food for thought. *Instr. Sci.* **38**, 105–134 (2010).

103. Messick, S. Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educ. Res.* **18**, 5–11 (1989).

104. Schuwirth, L. W. T. & Van Der Vleuten, C. P. M. General overview of the theories used in assessment: AMEE Guide No. 57. *Med. Teach.* **33**, 783–797 (2011).

105. Cook, D. A. & Beckman, T. J. Current concepts in validity and reliability for psychometric instruments: Theory and application. *American Journal of Medicine* (2006). doi:10.1016/j.amjmed.2005.10.036

106. Downing, S. M. Face validity of assessments: Faith-based interpretations or evidence-based science? *Med. Educ.* **40**, 7–8 (2006).

107. Korndorffer, J. R., Kasten, S. J. & Downing, S. M. Association for Surgical Education A call for the utilization of consensus standards in the surgical education literature. *AJS* **199**, 99–104 (2010).

108. Cook, D. A. & Hatala, R. Validation of educational assessments: a primer for simulation and beyond. *Adv. Simul.* **1**, 31 (2016).

109. Borgersen, N. J. *et al.* Gathering Validity Evidence for Surgical Simulation. *Ann. Surg.* 1 (2018). doi:10.1097/SLA.0000000000002652

110. Messick, S. Validity. (1987).

111. Ghaderi, I. *et al.* Technical skills assessment toolbox a review using the unitary framework of validity. *Annals of Surgery* **261**, 251–262 (2015).

112. American Psychological Association; National Council on Measurement in Education; Joint Committee on Standards for Educational and Psychological Testing, U. . *The Standards for Educational and Psychological Testing*.

113. Downing, S. M. Validity: On the meaningful interpretation of assessment data. *Med. Educ.* **37**, 830–837 (2003).

114. Cortina, J. M. What Is Coefficient Alpha? An Examination of Theory and Applications. *J. Appl. Psychol.* **78**, 98–104 (1993).

115. Downing, S. M. Reliability: On the reproducibility of assessment data. *Med. Educ.* **38**, 1006–1012 (2004).

116. Ebel, R. L. Estimation of the reliability of ratings. *Psychometrika* **16**, 407–424 (1951).

117. Bloch, R. & Norman, G. Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. *Med. Teach.* **34**, 960–992 (2012).

118. Mushquash, C. & O'Connor, B. P. SPSS and SAS programs for generalizability theory analyses. *Behav. Res. Methods* **38**, 542–7 (2006).

119. Ben-David, M. F. AMEE guide no. 18: Standard setting in student assessment. *Med. Teach.* **22**, 120–130 (2000).

120. Van Nijlen, D. & Janssen, R. *Modeling Judgments in the Angoff and Contrasting-Groups Method of Standard Setting. Journal of Educational Measurement Spring* **45**, (Kane, 2008).

121. Zieky, M. J. So Much Has Changed: How the Setting of Cutscores Has Evolved Since the 1980s. in *Setting Performance Standards* 33–66 (Routledge,

2001). doi:10.4324/9781410600264-7

122. Gustafsson, A. *et al.* Hip-fracture osteosynthesis training: exploring learning curves and setting proficiency standards. *Acta Orthop.* **90**, 348–353 (2019).

123. De Montbrun, S., Satterthwaite, L. & Grantcharov, T. P. Setting pass scores for assessment of technical performance by surgical trainees. (2014). doi:10.1002/bjs.10047

124. Setting Performance Standards on Tests. in *Handbook of Test Development* 228–254 (Routledge, 2015). doi:10.4324/9780203102961-18

125. Jørgensen, M., Konge, L. & Subhi, Y. Contrasting groups' standard setting for consequences analysis in validity studies: reporting considerations. *Adv. Simul.* **3**, 5 (2018).

126. Sethia, R., Kerwin, T. F. & Wiet, G. J. Performance Assessment for Mastoidectomy: State of the Art Review. *Otolaryngol. Neck Surg.* **156**, 61–69 (2017).

127. Butler, N. N. & Wiet, G. J. Reliability of the welling scale (WS1) for rating temporal bone dissection performance. *Laryngoscope* **117**, 1803–1808 (2007).

128. Balkany, T. *et al.* Development and Validation of the Cochlear Implant Surgical Competency Assessment Instrument. *Otol. Neurotol.* **38**, 504–509 (2017).

129. Piromchai, P. *et al.* The construct validity and reliability of an assessment tool for competency in cochlear implant surgery. *Biomed Res. Int.* **2014**, (2014).

130. Andersen, S. A. W., Cayé-Thomasen, P. & Sørensen, M. S. Mastoidectomy performance assessment of virtual simulation training using final-product analysis. *Laryngoscope* **125**, 431–435 (2015).

131. Messick, S. Validity. *ETS Res. Rep. Ser.* **1987**, i–208 (1987).

132. Martin, J. A. *et al.* Objective structured assessment of technical skill (OSATS) for surgical residents. *Br. J. Surg.* **84**, 273–278 (1997).

133. Frithioff, A., Sørensen, M. S. & Andersen, S. A. W. European status on temporal bone training: a questionnaire study. *Eur. Arch. Oto-Rhino-Laryngology* **275**, 357–363 (2018).

134. Andersen, S. A. W., Cayé-Thomasen, P. & Sørensen, M. S. Mastoidectomy performance assessment of virtual simulation training using final-product analysis. *Laryngoscope* **125**, 431–435 (2015).

135. Ilgen, J. S., Ma, I. W. Y., Hatala, R. & Cook, D. A. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Med. Educ.* **49**, 161–173 (2015).

136. Cook, D. A., Zendejas, B., Hamstra, S. J., Hatala, R. & Brydges, R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv. Heal. Sci. Educ.* **19**, 233–250 (2014).

137. Szasz, P., Louridas, M., Harris, K. A., Aggarwal, R. & Grantcharov, T. P. Assessing Technical Competence in Surgical Trainees. *Ann. Surg.* **261**, 1046–1055 (2015).

138. Downing, S. M. & Haladyna, T. M. Validity threats: Overcoming interference with proposed interpretations of assessment data. *Med. Educ.* **38**, 327–333 (2004).

139. Downing, S. M. Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education* **7**, 235–241 (2002).

140. Arora, A. *et al.* Face and content validation of a virtual reality temporal bone

simulator. *Otolaryngol. - Head Neck Surg.* **146**, 497–503 (2012).

141. Cook, D. A. Much ado about differences: why expert-novice comparisons add little to the validity argument. *Adv. Heal. Sci. Educ.* **20**, 829–834 (2015).

142. Talks, B. J. *et al.* The Melbourne Mastoidectomy Scale: validation of an end□product dissection scale for cortical mastoidectomy. *Clin. Otolaryngol.* (2020). doi:10.1111/coa.13569

143. Forslund Jacobsen, M. *et al.* Simulation of advanced cataract surgery – validation of a newly developed test. *Acta Ophthalmol.* **98**, 687–692 (2020).

144. Stefanidis, D., Acker, C. E., Swiderski, D., Heniford, B. T. & Greene, F. L. Challenges During the Implementation of a Laparoscopic Skills Curriculum in a Busy General Surgery Residency Program. *Journal of Surgical Education* **65**, 4–7 (2008).

145. Stefanidis, D., Scerbo, M. W., Montero, P. N., Acker, C. E. & Smith, W. D. Simulator training to automaticity leads to improved skill transfer compared with traditional proficiency-based training: A randomized controlled trial. *Ann. Surg.* **255**, 30–37 (2012).

146. Yudkowsky, R., Park, Y. S., Lineberry, M., Knox, A. & Ritter, E. M. Setting mastery learning standards. *Acad. Med.* **90**, 1495–1500 (2015).

147. Madsen, M. E. *et al.* Assessment of performance measures and learning curves for use of a virtual-reality ultrasound simulator in transvaginal ultrasound examination. *Ultrasound Obstet. Gynecol.* **44**, 693–699 (2014).

148. Walsh, C., Ling, S., Khanna, N., … M. C.-G. & 2014, undefined. Gastrointestinal Endoscopy Competency Assessment Tool: development of a procedure-specific assessment tool for colonoscopy. *Elsevier*

149. von Buchwald, J. H., Frendø, M., Guldager, M. J., Melchiors, J. & Andersen,

S. A. W. Content validity evidence for a simulation-based test of handheld otoscopy skills. *Eur. Arch. Oto-Rhino-Laryngology* (2020). doi:10.1007/s00405-020-06336-6

150. Fink, A., Kosecoff, J., Chassin, M. & Brook, R. H. Consensus methods: Characteristics and guidelines for use. *Am. J. Public Health* **74**, 979–983 (1984).

151. Jones, J. & Hunter, D. Consensus methods for medical and health services research. *BMJ* **311**, 376–80 (1995).

152. Humphrey-Murto, S., Varpio, L., Gonsalves, C. & Wood, T. J. Using consensus group methods such as Delphi and Nominal Group in medical education research. *Med. Teach.* **39**, 14–19 (2017).

153. Grantcharov, T. P. & Funch-Jensen, P. Can everyone achieve proficiency with the laparoscopic technique? Learning curve patterns in technical skills acquisition. *Am. J. Surg.* **197**, 447–449 (2009).

154. Roosli, C., Sim, J. H., Möckel, H., Mokosch, M. & Probst, R. An artificial temporal bone as a training tool for cochlear implantation. *Otol. Neurotol.* **34**, 1048–1051 (2013).

155. Frithioff, A., Frendø, M., Pedersen, D. B., Sørensen, M. S. & Andersen, S. A. W. 3D-printed models for temporal bone surgical training: A systematic review. *Otolaryngol. Neck Surg.* **Accepted**, (2021).

156. Guskey, T. R. Mastery Learning : Applying Theory. *Theory Pract.* **19**, 104–111 (1980).

157. Andersen, S. A. W., Konge, L., Cayé-Thomasen, P. & Sørensen, M. S. Learning curves of virtual mastoidectomy in distributed and massed practice. *JAMA Otolaryngol. - Head Neck Surg.* **141**, 913–918 (2015).

158. Nash, R. *et al.* Objective assessment of learning curves for the Voxel-Man TempoSurg temporal bone surgery computer simulator. *J. Laryngol. Otol.* **126**, 663–669 (2012).

159. Jun Tang, Songhua Tan, Qin Fang, Wenjie Miao & Anzhou Tang. [Investigate of the learning curve of cochlear implantation] - PubMed. *Zhonghua Er Bi Yan Hou Tou Jing Wai Ke Za Zhi* 649–53 (2014).

160. Cook, D. A., Brydges, R., Zendejas, B., Hamstra, S. J. & Hatala, R. Mastery learning for health professionals using technology-enhanced simulation: A systematic review and meta-analysis. *Acad. Med.* **88**, 1178–1186 (2013).

161. Grober, E. D. *et al.* The Educational Impact of Bench Model Fidelity on the Acquisition of Technical Skill The Use of Clinically Relevant Outcome Measures. *Ann. Surg.* **240**, 374–381 (2004).

162. Frithioff, A., Frendø, M., Mikkelsen, P. T., Sørensen, M. S. & Andersen, S. A. W. Ultra-high-fidelity virtual reality mastoidectomy simulation training: a randomized, controlled trial. *Eur. Arch. Oto-Rhino-Laryngology* **277**, 1335–1341 (2020).

163. Cook, D. A., Bordage, G. & Schmidt, H. G. Description, justification and clarification: A framework for classifying the purposes of research in medical education. *Med. Educ.* **42**, 128–133 (2008).

164. Tolsgaard, M. G. *et al.* The myth of ivory tower versus practice-oriented research: A systematic review of randomised studies in medical education. *Med. Educ.* 1–8 (2020). doi:10.1111/medu.14373

165. McGaghie, W. C., Issenberg, S. B., Petrusa, E. R. & Scalese, R. J. Revisiting 'A critical review of simulation-based medical education research: 2003–2009'. *Med. Educ.* **50**, 986–991 (2016).

166. Cook, D. A. If you teach them, they will learn: why medical education needs comparative effectiveness research. *Adv. Heal. Sci. Educ.* **17**, 305–310 (2012).

167. Cook, D. A. The research we still are not doing: An agenda for the study of computer-based learning. *Acad. Med.* **80**, 541–548 (2005).

168. Fjørtoft, K., Konge, L., Gögenur, I. & Thinggaard, E. The Implementation Gap in Laparoscopic Simulation Training. *Scand. J. Surg.* 145749691879820 (2018). doi:10.1177/1457496918798201

169. Thinggaard, E. *et al.* Take-home training in a simulation-based laparoscopy course. *Surg. Endosc. Other Interv. Tech.* **31**, 1738–1745 (2017).

170. Korndorffer, J. R., Bellows, C. F., Tekian, A., Harris, I. B. & Downing, S. M. Effective home laparoscopic simulation training: A preliminary evaluation of an improved training paradigm. *Am. J. Surg.* **203**, 1–7 (2012).

171. Wood, T. J. & Pugh, D. Are rating scales really better than checklists for measuring increasing levels of expertise? *Med. Teach.* **42**, 46–51 (2020).

172. Hodges, B., Regehr, G., McNaughton, N., Tiberius, R. & Hanson, M. OSCE checklists do not capture increasing levels of expertise. *Acad. Med.* **74**, 1129–1134 (1999).

173. Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **15**, 155–163 (2016).

174. Andersen, S. A. W., Park, Y. S., Sørensen, M. S. & Konge, L. Reliable Assessment of Surgical Technical Skills Is Dependent on Context: An Exploration of Different Variables Using Generalizability Theory. *Acad. Med.* **95**, 1929–1936 (2020).

175. Thomsen, A. S. S., Kiilgaard, J. F., Kjærbo, H., La Cour, M. & Konge, L.

Simulation-based certification for cataract surgery. *Acta Ophthalmol.* **93**, 416–421 (2015).

176. McKechnie, T. *et al.* Virtual Surgical Training During COVID-19: Operating Room Simulation Platforms Accessible From Home. *Ann. Surg.* **272**, e153–e154 (2020).

177. Liang, M. H. Longitudinal construct validity: Establishment of clinical meaning in patient evaluative instruments. *Med. Care* **38**, (2000).

178. Streiner, D. L., Norman, G. R. & Cairney, J. *Health Measurement Scales*. **1**, (Oxford University Press, 2015).

179. Arthur, W., Bennett, W., Stanush, P. L. & McNelly, T. L. Factors that influence skill decay and retention: A quantitative review and analysis. *Hum. Perform.* **11**, 57–101 (1998).

180. Ericsson, K. A. Deliberate practice and acquisition of expert performance: A general overview. *Acad. Emerg. Med.* **15**, 988–994 (2008).

181. Malik, M. U. *et al.* Determinants of resident competence in mastoidectomy: Role of interest and deliberate practice. *Laryngoscope* **123**, 3162–3167 (2013).

182. Choi, W. J., Jung, J. J. & Grantcharov, T. P. Impact of hawthorne effect on healthcare professionals: A systematic review. *Univ. Toronto Med. J.* **96**, 21–32 (2019).

183. Cook, D. A. & Beckman, T. J. Reflections on experimental research in medical education. *Adv. Heal. Sci. Educ.* **15**, 455–464 (2010).

184. Brighton, B., Bhandari, M., Tornetta, P. & Felson, D. T. Hierarchy of Evidence: From Case Reports to Randomized Controlled Trials. *Clin. Orthop. Relat. Res.* **413**, 19–24 (2003).

185. Norman, G. Data dredging, salami-slicing, and other successful strategies to

ensure rejection: Twelve tips on how to not get your paper published. *Adv. Heal. Sci. Educ.* **19**, 1–5 (2014).

186. Trognon, A. M. & Batt, M. Group Dynamics and Learning. in *Encyclopedia of the Sciences of Learning* 1388–1391 (Springer US, 2012). doi:10.1007/978-1-4419-1428-6_1863