Published in final edited form as:

Otolaryngol Head Neck Surg. 2017 January ; 156(1): 61–69. doi:10.1177/0194599816670886.

Performance assessment for mastoidectomy: State of the Art Review

Rishabh Sethia, BS¹, Thomas F. Kerwin, PhD², and Gregory J. Wiet, MD, FAAP, FACS^{3,4}

¹College of Medicine, The Ohio State University, Columbus, OH, USA

²Ohio Supercomputer Center, Columbus, OH, USA

³Department of Otolaryngology, The Ohio State University, Columbus, OH, USA

⁴Department of Otolaryngology, Nationwide Children's Hospital, Columbus, OH, USA

Abstract

Objective—The aim of this report is to provide a review of the current literature for assessment of performance for mastoidectomy, to identify the current assessment tools available in the literature, and to summarize the evidence for their validity.

Data Sources—The MEDLINE database was accessed via PubMed.

Review Methods—Inclusion criteria consisted of English-language published articles that reported use of a mastoidectomy performance assessment tool. Studies ranged from 2007 to November 2015, and were divided into two groups: intra-operative assessments or those performed with simulation (cadaveric laboratory or virtual reality). Studies that contained specific reliability analyses were also highlighted. For each publication, validity evidence data was analyzed and interpreted based on conceptual definitions provided in a recent systematic review on the modern framework of validity evidence.

Conclusions—Twenty-three studies were identified that met our inclusion criteria for review including four intra-operative objective assessment studies, five cadaveric studies, ten virtual reality (VR) simulation studies, and four that used both cadaveric and VR.

Implications for Practice—A review of the literature revealed a wide variety of mastoidectomy assessment tools and varying levels of reliability and validity evidence. The assessment tool developed at Johns Hopkins possesses the most validity evidence of those reviewed. However, there are a number of agreed upon specific metrics which could be integrated into a standardized assessment instrument to be used nationally. A universally agreed upon assessment tool will provide a means for developing standardized benchmarks for performing mastoid surgery.

Keywords

objective assessment; operative competency; technical performance; mastoidectomy; validity evidence

Introduction

In recent years, competence assessment has become an integral part of the medical field. Historically, surgical training programs have struggled with assessment of technical performance due to the use of highly subjective and poorly standardized and validated evaluation techniques. In response, a growing area of research has emerged with the goal of developing and validating objective assessment tools to measure technical skill in surgical trainees.¹ Within Otolaryngology, and more specifically in the field of Otology, researchers have largely focused on demonstrating validity and reliability evidence for these newly developed assessment tools for mastoidectomy. The following review will highlight and evaluate the current tools used by residency programs to assess mastoidectomy performance in the operating room (OR), cadaveric laboratory, or with virtual reality (VR) simulation.

In order to further examine competency assessment, we must first address a key question: what defines surgical competence? According to Bhatti et al., a competent surgeon must successfully apply professional knowledge, skills, and attitude to new situations as well as to familiar tasks.² Furthermore, assessment of technical performance must include evaluation of a surgeon's dexterity, knowledge of the procedure, and decision-making during the procedure.¹ Over the years, many researchers have attempted to create an ideal tool that incorporates all aspects of surgical competence in its assessment. However, surgical residency programs currently use a variety of tools to assess trainees, and no single tool has emerged as the "gold-standard". The time honored assessment currently used by the American Board of Otolaryngology (ABOto) and the Accreditation Council for Graduate Medical Education (ACGME) is based upon accumulation of "adequate" case numbers during training and the attestation of the specific residency program director where the resident trained as to the resident's technical skill as well as written and oral examinations.^{3,4}

Recently, residency programs have shifted towards a formal competency-based curriculum to overcome increased workload, work-hour restrictions, and a history of subjective assessment, which has prevented adequate evaluation of resident competency.⁵ In July 2001, the ACGME introduced the Outcome Project in order to address some of these concerns through the widespread integration of a standardized, competency-based resident assessment program. In order to maintain full accreditation, programs must utilize and document the ACGME's specialty-specific Milestones to formally assess resident competence.³ Although the ACGME has standardized the process, a recent review of 1,076 Otolaryngology residency program directors surveyed from 2001 to 2012 revealed that a lack of financial support, protected time, and personnel has limited the implementation of these new mandates.⁶ In order to combat these barriers, program directors have suggested creating a "practical clearinghouse" of existing and emerging assessment instruments.⁷ Hence, new research has focused on the development of time- and resource-efficient assessment tools.

In addition to being efficient, a surgical assessment instrument must contain certain key characteristics in order to become widely accepted. The tool must be reliable, feasible, fair, objective, and valid.⁸ Validity refers to whether or not an assessment provides an accurate measure of what it is intended to evaluate. Reliability refers to the consistency of the

assessment and asks the question: is this assessment reproducible? For example, a scale that displays the same weight every time but underestimates the weight significantly can be considered reliable, but contains poor validity.

The scope of an assessment should allow the surgical training program to evaluate the resident's ability to handle all job-related tasks.⁹ Within Otolaryngology, to the best of our knowledge, a standardized, objective tool that contains all of the elements necessary for proper assessment has not been formally recognized by the ACGME and/or the ABOto for mastoidectomy evaluation, although an attempt to create a cross-institutional scale has been reported in the literature.¹⁰ For this reason, we provide the following overview of the various tools currently in use, and we evaluate their strengths and limitations.

Current Assessment Tools

One of the most widely recognized surgical assessment tools reported in the literature is the Objective Structured Assessment of Technical Skill (OSATS). First developed by Martin et al. at the University of Toronto, the OSATS has been shown to be a reliable and valid framework for developing procedure-specific operative skills in surgical trainees for both bench model and live animal simulations.¹¹ This objective examination consists of two components: a Global Rating Scale (GRS), a Task-Based Checklist (TBC). The GRS and TBC offer a more detailed, objective assessment of technical skill compared to a single global rating by a surgical resident's preceptor. The TBC was tailored to the specific procedure and contained between 22–32 steps for assessment. Each item was scored with 1 point if done correctly, and 0 if not done or done incorrectly. The GRS consisted of a 5-point scale and was not task-specific, but rather a holistic assessment of surgical performance. It includes aspects present throughout the procedure such as respect for tissue, time and motion, instrument handling, use of assistants, flow of operation and forward planning, and knowledge of the instruments and procedure. Variations of these individual components have since been widely disseminated and applied in the field of mastoidectomy assessment.

Martin's paper demonstrated use of the OSATS tool for a very generalized purpose in assessing general surgical skill. The instrument measured performance in excision of skin lesions, handsewn bowel anastomosis, stapled bowel anastomosis, insertion of a T tube, abdominal wall closure and control of inferior vena cava haemorrhage. In 1997, the same group further illustrated interstation reliability and construct validity for the OSATS on a larger scale, and determined that the GRS was slightly more reliable than the TBC.¹² Reznick's paper added pyloroplasty and tracheostomy procedures to that originally described by Martin.

Since the introduction of this initial objective assessment framework for general surgical trainees, researchers in Otolaryngology have introduced several mastoidectomy assessment instruments that incorporate the same methodology. Task-specific checklists and global rating specific to mastoidectomy were developed. The first example in the literature of this was Zirkle et al. from the University of Toronto who incorporated TBC, GRS, as well as Final Product Analysis (FPA) and Expert Opinion (EO).¹³ The authors defined FPA as a measure of the quality of the end product only, rather than evaluation of the process, such as

in TBC or GRS. EO was defined as the standard of assessment at the time (subjective assessment by preceptor) and was compared to scores from the GRS, TBC, and FPA.

Several more scales were developed by other groups after the Zirkle et al. study. A group at Johns Hopkins developed an OSATS for mastoidectomy performance containing both a TBC and GRS.^{14,15} The Welling Scale (WS1) developed at The Ohio State University (OSU) measures FPA for a complete mastoidectomy with facial recess, while a newer report from Mowry et al. introduced the Iowa Standardized Skill Assessment, which also uses FPA.^{16,17} Piromchai et al. developed a related OSATS for cochlear implant surgery incorporating only two particular TBC components related to mastoidectomy.¹⁸ Several VR-based mastoidectomy assessments have been reported in the literature as well.^{19,20} The diversity of evaluation methods noted above is probably due to the status of the field of temporal bone performance assessment in general. When these studies were published, there was no consensus on what the best method was for assessment. Use of the OSATS requires simultaneous observation of the procedure (either in live surgery or video replay) and is particularly time-consuming. We postulate that the time to perform an FPA is significantly shorter and therefore more appealing to expert raters. This certainly was the case with our own study using cadaveric temporal bones¹⁶.

Methods

Using PubMed to access the MEDLINE database, a search for studies that reported use of mastoidectomy assessment was performed through November 2015. Search terms included “mastoidectomy performance assessment,” and “mastoidectomy competence”. Additional references were obtained from the citations of the articles obtained through the MEDLINE search. Only articles published in the English-language that reported use of a mastoidectomy performance assessment tool were included. Studies ranged from 2007 to 2015, and were divided into two groups: intra-operative assessments or those performed with simulation (cadaveric laboratory or virtual reality).

Validity Evidence

In a recent systematic review on the validity of simulation-based assessments, Cook et al.²¹ described a modern framework consisting of five main sources of validity evidence that was originally adopted as standard by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) in 1999.²² In this model, all validity is considered construct validity. Evidence sources include content, response process, internal structure, relations with other variables, and consequences (Table 3).

The following brief definitions of the validity evidence sources were transcribed from the 2013 systematic review by Cook et al.²¹. Please note that these definitions have been condensed and are intended to provide the reader with a general overview of validity evidence. More detailed definitions can be found in the referenced article, as well as the original description of the framework.^{21,22}

Content evidence—A description of steps taken to ensure that assessment content reflects the construct it is intended to measure.

Response process evidence—The theoretical and empirical analyses evaluating how well rater or examinee actions align with the intended construct.

Internal structure evidence—The data evaluating relations among individual assessment items and how these relate to the overarching construct.

Relations with other variable evidence—The statistical associations between assessment scores and another measure or feature that has a specified theoretical relationship.

Consequences evidence—The impact, beneficial or harmful, of the assessment itself and the decisions and actions that result.

Although the Cook study focused on simulation-based assessments, the authors noted that the structure and definitions provided can and should be applied broadly to all educational assessments.²¹ Using these conceptual definitions, we evaluated the level of validity evidence presented in each study as seen in Tables 1 and 2. In this framework, evidence such as inter-item, inter-station, and inter-rater reliability are included within internal structure evidence. As such, we performed a separate analysis of reliability (Table 4). Of note, Tables 1 and 2 list the new validity evidence presented by each individual study rather than list any evidence that has been accumulated for the instrument since its introduction.

Assessment Studies

Twenty-three studies were identified in total. Four involved live assessments in the OR (Table 1). These were all based on the Hopkins tool. Nineteen were simulation-based. Of the studies involving simulation, five used cadaveric laboratory temporal bone dissection (TBD), ten used VR simulation, and four used both (Table 2). Two cadaveric assessment studies used the Hopkins scale, while the remaining three used the Iowa, WS1, and a modified Zirkle¹³ OSATS scale.

A majority of the VR assessment studies focused on the validation of automated feedback and assessment by comparing algorithm-derived internal performance metrics from the VR simulator to EO, or to a modified, validated scale such as the WS1 or a modified Hopkins OSATS.^{15,19,20,23,24,25} The VR systems employed in these studies include the Stanford, OSU, University of Melbourne, Visible Ear, and VOXEL-MAN simulators.

Sewell et al. validated automated scoring metrics for the mastoidectomy simulator at Stanford and established reliability between computed scores and expert global rating scores.¹⁹ Kerwin et al. at OSU performed a similar validation of automated metrics by comparing student-drilled temporal bones to expert-drilled bones.²⁴ At University of Melbourne, Wijewickrema et al. showed that their simulator provided timely and appropriate feedback according to postexperiment expert observation.²⁰ Khemani et al. demonstrated “construct validity” for objective metrics derived from the VOXEL-MAN simulator,²³ while

Francis et al. also used the VOXEL-MAN to show that technical skills improved after practice on the simulator using their modified Hopkins OSATS.¹⁵

All simulation-based studies presented some type of relations with other variables evidence, with the exception of the Wan et al. study. In addition, many reported internal structure and content evidence, with no reports of consequences or response process evidence (Tables 1 & 2). The majority of studies reported inter-rater reliability using various statistical analyses and utilized 2–3 (often including the computer/VR simulator itself) raters from a single institution. However, few studies reported per-item reliability analysis (Table 4).

Discussion

Nearly all studies presented in this review reported using an older concept of “construct validity.” This older concept uses differences in scores between novices and experts as evidence that the assessment is valid. Using a more modern framework, this type of evidence would fit under the definition of relations with other variables evidence. However, Cook et al. concluded that this novice/expert differentiation constitutes only weak validity evidence. They claim that by this standard, gray hair and baldness would be sufficient to discriminate between senior surgeons and younger medical students.²¹ In addition, many studies that qualified for internal structure evidence reported only inter-rater reliability, with only six studies^{14,15,16,19,24,26} also reporting per-item reliability (Table 4). Unique to the WS1, Butler et al. demonstrated both inter- and intra-rater reliability.¹⁶ Overall, the validity evidence summarized in this review corresponds with previous reports in that technical skill evaluation using observational and simulation-based assessment tools is currently not reliable and valid at the specialist level.^{21,27,28}

Assessment Instruments

All four of the OR assessments reported in the literature were published by the group at Johns Hopkins using their tool. The Hopkins scale, which incorporates a TBC and GRS, was validated for use in the OR by Francis et al. in 2010 and has since been reported in numerous studies covering different aspects of mastoidectomy assessment.²⁶ For example, one study determined the average time required to complete a mastoidectomy assessment was 7.5 days following completion of the procedure, and that assessment tool “construct validity” was lost if completed after 6 days.²⁹ This emphasized the importance of completing the assessment in a timely fashion from when the observation was made. Another discovered that an interest in otology and cumulative number of cases were positively associated with mastoidectomy performance, while self-directed laboratory simulation practice was negatively associated.³⁰ Finally, the group also developed three sets of technical milestones based on stages of their TBC to establish training benchmarks. These were based on the number of cases to achieve competency at each milestone and intended to better identify residents in need of remediation using this scale.³¹ All four OR studies exhibited evidence of relations with other variables, while two also contained internal structure evidence, and one demonstrated consequences evidence.

Laeq et al. first introduced the Hopkins scale in 2009 by demonstrating its feasibility, validity, and reliability using cadaveric assessment.¹⁴ Since its introduction, the Hopkins

scale has accumulated significant validity evidence in various settings and by multiple groups. Recently, Awad et al. demonstrated construct validity in the cadaveric lab as the first “outside” group to use the Hopkins instrument.³² Therefore, the Hopkins scale likely contains the most validity evidence for mastoidectomy performance assessment use of the tools described.

In 2007, Butler and Wiet demonstrated very good to excellent inter- and intra-rater reliability for the Welling Scale.¹⁶ The following year, Fernandez et al. demonstrated the validity of the WS1 using generalizability theory.³³ A modified WS1 was also recently used in several recent reports using the Visible Ear Simulator. West et al.³⁴ demonstrated peak and ceiling limitations of FPA, while Andersen et al. determined that self-directed VR simulation increased cadaveric dissection performance and that distributed practice was superior to massed practice with regard to cognitive load and learning curves.^{35,36,37} In these studies, reaction time to activate a button on the simulator was used as a surrogate to cognitive load. Those individuals randomized to distributed practice had significantly better reaction times implying a lower overall cognitive load while performing the mastoidectomy in their simulator. In 2011, Zhao et al. used both the WS1 and a custom OSATS derived from previous instruments to show that VR training results in improved performance on cadaveric TBD assessment.³⁸ Recently, Andersen et al. validated the metrics of an automated VR assessment on the Visible Ear Simulator by using scores obtained from WS1 evaluations of cadaveric TBD.³⁹ Finally, Andersen et al. also validated the use of a modified WS1 for VR assessment with the Visible Ear simulator.³⁵ Overall, the Welling Scale has accumulated significant validity evidence in the literature by multiple groups, similar to the Hopkins scale. Furthermore, the scale provides the unique assessment method of final product analysis in comparison to the Hopkins scale.

While the Iowa Standardized Skill Assessment and WS1 both used FPA only, little has been reported on the Iowa instrument outside of one retrospective study involving cadaveric assessment.¹⁷ Of the instruments reported in the literature, the Iowa instrument has accumulated the least validity evidence limiting its use in performance assessment.

As mentioned earlier, Zirkle et al. validated the expert opinion, global rating scale, and task-based checklist components of their OSATS, and showed TBC correlated best with EO while FPA did not correlate with trainee experience.¹³ In 2007, Zirkle et al. demonstrated construct validity for hand motion analysis using both VR and cadaveric temporal bone dissection.⁴⁰ The instrument introduced by Zirkle et al. is the only one reported that contains GRS, TBC, FPA, and EO while exhibiting content, internal structure, relations with other variables, and consequences evidence. Since these initial reports, however, further validation studies for the instrument have not been widely reported in contrast to the Hopkins and Welling scales.

Finally, a study by Wan et al. from OSU combined aspects of the WS1 scale, The University of Toronto scale, and Stanford metrics in an attempt to standardize objective assessments. Members of the American Neurotology Society and American Otological Society were surveyed to rank the importance of the criteria, and those items ranked by >70% of respondents as “very important” or “important” were used to propose a 24-item cross-

institutional scale. Both task-specific and global rating components were incorporated.¹⁰ As this scale has only been reported in one study, it serves more as a model for an ideal scale based on a combination of all of the instruments currently employed in the field, though it too has its limitations.

Validity Issues

It is important to remember that OSATS was created as a framework for developing assessment tools. The fact that the OSATS framework was used to create a tool does not endow the tool with validity or reliability. It has been shown to be a useful strategy for developing tools, but authors should refrain from implying validity of a specific tool based on the use of the OSATS strategy.

With regards to evaluating reliability of assessment tools, a common analysis performed is the inter-rater reliability of the entire tool. The individual items in the FPA or TBC are summed, and that sum is measured to give an overall score. This is a necessary quality for an objective test tool. However, another important area for analysis that is often overlooked is inter-rater reliability for the individual items on the tool. If the items reflect different types of skills and the reliability and validity for the individual items on the tool are not validated, you cannot assess the areas that a test-taker has failed. Although good for summative processes, using an overall score dramatically limits the usefulness of the result in providing formative feedback to the test-taker. Item-specific analysis makes the overall process more useful for detection of specific strengths or weaknesses.

The requirement of inter-rater reliability for individual items is especially important if the further goal is to create automatic assessment systems. Automatic assessment systems require a standard to compare their results against. If there is no agreed standard for an individual item on a test, an automated system cannot provide feedback for where a particular deficiency lies and what needs to be done to improve.

Furthermore, in order to create tests that are valid cross-institutionally, the rating data must be acquired cross-institutionally. By in large, this has not been the case with the studies listed. They have primarily been developed and deployed at a single institution (Table 4). A single institution potentially provides consistent understanding of the definition of individual metrics, but studies done at one institution do little to promote dissemination. In our experience performing a multi-institutional study, interpretations of complex metrics can vary a great deal. In a recent preliminary study, we had experienced surgeons from 13 institutions around the United States act as raters of virtual mastoidectomies. Despite using a Delphi method to come to agreement on the definitions of each individual item, kappa scores demonstrated at best only moderate agreement. Some of the inconsistencies in the rating data were explained during follow up face-to-face discussion in which different surgeons disagreed on didactic methods, best practices, and interpretations of specific metrics even though they had previously agreed to definitions of each item. It was necessary to rewrite the specific metrics on the instrument to make them more precise and to obtain higher consensus between surgeons with different backgrounds. The training program of raters may affect the decisions that they make in grading items on an assessment tool, as does their current institutional practice. Without other information, we assume that the raters in the studies

discussed here are only from the study authors' institutions. To strengthen a claim of inter-rater reliability and improve overall validity, it is desirable to have a large group of raters from a large variety of institutions and training backgrounds train on and administer a rating instrument.

Finally, we include a comparison of the task-based scales presented in this study (Appendix 1). Tasks listed under the "Cross-Institutional" scale represent criteria ranked as "very important" or "important" by >70% of experts as mentioned previously.¹⁰ The tasks were divided into ten major components of mastoidectomy assessment. The scales address different aspects of the procedure with most TBC's missing several major components, further highlighting the variability that exists in the field. The cross-institutional scale by Wan et al. omits the digastric component, while the Hopkins and Iowa scales omit both drill handling, and the semicircular canal and digastric components, respectively. Although the Hopkins scale lacks a drill handling component in their TBC, a nonspecific component is contained within their GRS.

In summary, the Hopkins scale has been most widely validated in different settings and by multiple institutions, while the Iowa scale has only been validated in one study. Use of the cross-institutional scale has not been reported in the literature since its introduction. Although the Welling scale has also been widely validated, the scale itself lacks important aspects such as antrum, ossicles/middle ear, or facial recess components and does not take into consideration the process of the surgery. Thus, we conclude that the Hopkins scale should be considered the most comprehensive of currently used mastoidectomy assessment tools, although dissemination and testing of this instrument to multiple institutions would bolster validity. An ideal tool would contain all ten components listed in the appendix that are common and be validated by multiple institutions using effective forms of validity and reliability evidence as proposed in this review.

Implications for Practice

In order for governing entities (ACGME and ABOto) to adopt these assessment tools on a national level, the evaluation method must contain strong validity and reliability evidence, and must be feasible to implement. This is especially important for time-constrained attending physicians tasked with making the assessments. Using the time limit proposed in the 2009 validation study by Laeeq et al.¹⁴, one can estimate an average of 35 minutes required per mastoidectomy assessment. In this study, evaluators conducted a total of 118 assessments over 3 years, yielding an average of 23 hours per year of expert time dedicated to mastoidectomy assessment. We consider this a conservative estimate given that Williams et al. recommended 38 observations for each resident per year to obtain reliable performance ratings (reliability score of 0.80).⁴¹ Recognizing this time commitment is important when instituting these types of expert observational assessments.

In recent years, researchers have developed algorithms for automated assessment by VR simulators driven by a pressure on our healthcare professional training system to provide efficient/cost-effective care in the midst of training. Assessment with feedback is known to be imperative to efficient and successful training, but measures of performance need to be

established in order to provide this feedback. Therefore, automated systems like those undergoing development using a multi-center network of institutions for testing and validation are important, especially those that make use of accepted metrics for performance measurement as the basis of feedback to the learner. Additional use of an objective automated assessment system could significantly reduce the time investment by attending physicians when administering an assessment tool themselves. It can also provide a level playing field with respect to difficulty of cases, objectivity, and convenience of administration of a professional objective technical skills assessment for board certification and continuance of practice.

In order to overcome the limitations of our current assessment tools, large multi-center trials of these tools, comparison with other testing mechanisms, following patient outcomes such as rates of re-operation, hearing outcomes, complication rates etc., and correlating this data with performance scores is necessary and will contribute to the acceptance of these tools. It is a very long process that requires organization and cooperation on the highest levels including the ACGME, ABOto, and government healthcare monitoring bodies.

Limitations

The validity evidence described in this review was determined by examining the methods of each study and subjectively assigning an appropriate level of evidence. Inconsistencies regarding the nature of what constitutes validity evidence are prevalent in the field. As such, we attempted to limit our biases by adhering to the conceptual definitions provided in the Cook systematic review on validity evidence.²¹

Conclusion

As residency programs have shifted to a competency-based training system, the validation of performance assessment tools has become a significant focus of research. The large number of instruments currently in use for mastoidectomy assessment highlights the need for a standardized, objective assessment tool to be implemented by Otolaryngology training programs. Each scale we analyzed lacks at least one of the ten widely accepted components of mastoidectomy assessment. Although the Hopkins scale should be considered the current leader in the field, work still needs to be done to increase the validity evidence for this tool before it can be universally accepted. Additionally, the time investment needed to administer this tool by experts needs to be considered and continued research into automated assessment tools such as that potentially provided in VR-type systems is necessary. We hope that this review will provide the framework for the systematic development of a widely accepted mastoidectomy assessment tool.

Acknowledgments

This work was supported by NIH/NIDCD R01-DC011321 (Wiet).

Appendix 1: Comparison of task-based scales

Tasks for the individual scales are listed. Each task was categorized into one of ten major components of mastoidectomy assessment as determined by the authors. None of the scales

contains all ten components. However, the Cross-Institutional, Hopkins, and Iowa scales omit fewer components than the Welling and Modified OSATS scales.

WSI (Dutler 2007)	Modified OSATS (Zirke 2007)	Murphy (Lacey 2009)	Cross-Institutional (Wen 2010)	Leiva (Munoz 2014)
Cortex	1. Selects appropriate cutting drill bit	1. Initial bone cuts	Maintains visibility while removing bone	Antrum defined
Cortex rounded at linea temporalis	2. Drills with broad strokes	a. Placement of superior cut	Selects appropriate burr type and size	Aditus opened/Incus defined
Cortex rounded from linea temporalis to middle cranial fossa	3. Drills with unidirectional strokes	b. Placement of canal cut	Antrum entered	Facial recess opened
Thinning of posterior canal wall	4. Dissection initiated within the anatomic confines	2. Defining anatomic limits	No violation of facial nerve sheath	Residual Air Cells
Complete sacralization	5. Dissection to the limits of a cortical mastoid	a. Identification and definition of tegmen	Sigmoid sinus not entered	Zygomatic Root
Tegmen mastoid	6. Exposes tegmen	b. Sharpen posterior EAC cortex	Identifies tympanic segment of the facial nerve	Sinudural Angle
Dissection parallels curve of the dura	7. Tegmen is not entered	c. Define sigmoid sinus and sino-dural angle	Does not drill on ossicle	Mastoid Tip
Completely exposed	8. Exposes sigmoid sinus	3. Open antrum	Does not use excessive drill force near critical structures	Tegmen Violation (with torn dura)
No holes	9. Sigmoid sinus is not entered	a. Deepen dissection at sino-dural angle	Firm, low, good hand position and grip on drill	<4mm
No cells remain	10. Defines sino-dural angle	b. Open antrum from posterior to anterior	Identifies the chorda tympani or stump	>4mm
Sigmoid Sinus	11. Thin posterior canal wall	c. Reveal short process of incus (traumatic)	Drills with broad strokes	Posterior Canal Wall Lowering
No holes	12. Does not enter posterior canal	4. Digastric dissection	Drills in best direction (clear understanding of cutting edge)	Any amount
No cells	13. Enters the antrum	a. Define cephalic edge digastric muscle	Identifies the facial nerve at the cochle-airiform process	Posterior Canal Wall Violation
No overhang	14. Does not drill on ossicles	b. Follow to SMF	Appropriate depth of cavity (cortex)	<4mm
Sinudural Angle	15. Edges of dissection rounded	5. Thin posterior EAC cortex (translucent)	Canal wall up (EAC)	>4mm
Sharp	16. Gradual slope of cavity	a. View posterior EAC on face	No holes in the EAC	Labyrinthine Damage
No cells remaining		b. Use side/front of appropriate burr	Complete sacralization (cortex)	Drilled
Digastric Ridge	Identified	c. Sacralization	Posterior canal wall thinned	Opened
Digastric tendon followed to stylomastoid foramen		6. Open facial recess	Low frequency of drill 'jumps' (jump' defined as drilling for	Ossicular Damage
External Auditory Canal		a. Even removal of infra-labyrinthine bone	Identifies facial nerve at the external genu	Incus drilled
Canal Wall Up		b. Medial thinning of EAC/cortex	Facial recess completely exposed (overlying bone sufficiently	Chain disruption
Without holes		c. Identify and preserve VII and chorda	No holes in tegmen	Sigmoid Inners
Without cells		d. Identify ME anatomy	Use of diamond burr within 2mm of facial nerve	Not defined
Semicircular Canals Skeletonized		e. Decompress facial nerve to SMF	No cells remain on sinudural angle	Opened
Horizontal		7. Posterior atticotomy		Facial Recess
Superior		a. Thin superior EAC cortex		No bottom
Posterior		b. Thin anterior tegmen		Chorda Violated
Blue lined without fenestra		c. Remove intervening bone		R/W exposure inadequate
Facial Nerve				Facial Nerve
Identification of nerve at the stylomastoid foramen				Uncovered
Identification of nerve at the external genu				Partial Tear
Identification tympanic segment				Disrupted
Identification of nerve at cochleariform process				
No exposed nerve sheath				
Identification of chorda tympani or stump				
Facial Recess Completely exposed				
Additional Anatomical Structures				
Stapedial muscle dissected				
EUS transition to duct				
Blue line of basal turn of the cochlea				
Identification of carotid artery in middle ear				
Identification of jugular bulb				
Skeletonization of jugular bulb				

References

1. Darzi A, Mackay S. Assessment of surgical competence. *Qual Health Care*. 2001;11:64–9. [PubMed: 11700382]
2. Bhatti NI, Cummings CW. Viewpoint: competency in surgical residency training: defining and raising the bar. *Acad Med*. 2007; 82(6):569–573. [PubMed: 17525542]
3. Accreditation Council for Graduate Medical Education. Aug 30. 2015 <https://www.acgme.org/acgmeweb>
4. American Board of Otolaryngology. Sep 3. 2015 <http://www.aboto.org>
5. Bhatti NI, Ahmed A. Improving skills development in residency using a deliberate-practice and learner- centered model. *Laryngoscope*. 2015; 125(S8):S1–S14.
6. Malik MU, Diaz Voss Varela DA, Stewart CM, et al. Barriers to implementing the ACGME outcome project: a systematic review of program director surveys. *J Grad Med Educ*. 2012; 4(4):425–433. [PubMed: 24294417]
7. Laeeq K, Weatherly RA, Masood H, et al. Barriers to the implementation of competency-based education and assessment: A survey of otolaryngology program directors. *Laryngoscope*. 2010; 120(6):1152–1158. [PubMed: 20513032]
8. Shah J, Darzi A. Surgical skills assessment: an ongoing debate. *BJU Int*. 2001; 88(7):655–660. [PubMed: 11890231]
9. Sidhu RS, Grober ED, Musselman LJ, Reznick RK. Assessing competency in surgery: where to begin? *Surgery*. 2004; 135(1):6–20. [PubMed: 14694296]
10. Wan D, Wiet GJ, Welling DB, Kerwin T, Stredney D. Creating a cross-institutional grading scale for temporal bone dissection. *Laryngoscope*. 2010; 120(7):1422–1427. [PubMed: 20564671]
11. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997; 84(2):273–278. [PubMed: 9052454]

12. Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative “bench station” examination. *Am J Surg.* 1997; 173(3):226–230. [PubMed: 9124632]
13. Zirkle M, Taplin MA, Anthony R, Dubrowski A. Objective assessment of temporal bone drilling skills. *Ann Otol Rhinol Laryngol.* 2007; 116(11):793–798. [PubMed: 18074662]
14. Laeeq K, Bhatti NI, Carey JP, et al. Pilot testing of an assessment tool for competency in mastoidectomy. *Laryngoscope.* 2009; 119(12):2402–2410. [PubMed: 19885831]
15. Francis HW, Malik MU, Diaz Voss Varela DA, et al. Technical skills improve after practice on virtual-reality temporal bone simulator. *Laryngoscope.* 2012; 122(6):1385–1391. [PubMed: 22487895]
16. Butler NN, Wiet GJ. Reliability of the Welling scale (WS1) for rating temporal bone dissection performance. *Laryngoscope.* 2007; 117(10):1803–1808. [PubMed: 17721407]
17. Mowry SE, Hansen MR. Resident participation in cadaveric temporal bone dissection correlates with improved performance on a standardized skill assessment instrument. *Otol Neurotol.* 2014; 35(1):77–83. [PubMed: 23988999]
18. Piromchai P, Kasemsiri P, Wijewickrema S, Ioannou I, Kennedy G, O’Leary S. The construct validity and reliability of an assessment tool for competency in cochlear implant surgery. *Biomed Res Int.* 2014:192741. [PubMed: 25114897]
19. Sewell C, Morris D, Blevins NH, et al. Validating metrics for a mastoidectomy simulator. *Stud Health Technol Inform.* 2007; 125:421–426. [PubMed: 17377316]
20. Wijewickrema S, Piromchai P, Zhou Y, et al. Developing Effective Automated Feedback in Temporal Bone Surgery Simulation. *Otolaryngol Head Neck Surg.* 2015 0194599815570880.
21. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Health Sci Educ.* 2013; 19(2):233–250.
22. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association; 1999.
23. Khemani S, Arora A, Singh A, Tolley N, Darzi A. Objective skills assessment and construct validation of a virtual reality temporal bone simulator. *Otol Neurotol.* 2012; 33(7):1225–1231. [PubMed: 22858711]
24. Kerwin T, Wiet G, Stredney D, Shen HW. Automatic scoring of virtual mastoidectomies using expert examples. *Int J Comput Assist Radiol Surg.* 2012; 7(1):1–11.
25. Linke R, Leichtle A, Sheikh F, et al. Assessment of skills using a virtual reality temporal bone surgery simulator. *Acta Otorhinolaryngol Ital.* 2013; 33:273–281. [PubMed: 24043916]
26. Francis HW, Masood H, Chaudhry KN, et al. Objective assessment of mastoidectomy skills in the operating room. *Otol Neurotol.* 2010; 31(5):759–765. [PubMed: 20517169]
27. Ahmed K, Miskovic D, Darzi A, Athanasiou T, Hanna GB. Observational tools for assessment of procedural skills: a systematic review. *Am J Surg.* 2011; 202(4):469–480. [PubMed: 21798511]
28. Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-enhanced simulation to assess health professionals: A systematic review of validity evidence, research methods, and reporting quality. *Acad Med.* 2013; 88(6):872–883. [PubMed: 23619073]
29. Laeeq K, Francis HW, Varela DA, Malik MU, Cummings CW, Bhatti NI. The timely completion of objective assessment tools for evaluation of technical skills. *Laryngoscope.* 2012; 122(11):2418–2421. [PubMed: 22865665]
30. Malik MU, Diaz Voss Varela DA, Park E, et al. Determinants of resident competence in mastoidectomy: role of interest and deliberate practice. *Laryngoscope.* 2013; 123(12):3162–3167. [PubMed: 23878112]
31. Francis HW, Masood H, Laeeq K, Bhatti NI. Defining milestones toward competency in mastoidectomy using a skills assessment paradigm. *Laryngoscope.* 2010; 120(7):1417–1421. [PubMed: 20578231]
32. Awad Z, Tornari C, Ahmed S, Tolley NS. Construct validity of cadaveric temporal bones for training and assessment in mastoidectomy. *Laryngoscope.* 2015; 125(10):2376–2381. [PubMed: 25891034]

33. Fernandez SA, Wiet GJ, Butler NN, Welling B, Jarjoura D. Reliability of surgical skills scores in otolaryngology residents: analysis using generalizability theory. *Eval Health Prof.* 2008; 31(4): 419–36. [PubMed: 18842619]
34. West N, Konge L, Cayé-Thomasen P, Sørensen MS, Andersen SA. Peak and ceiling effects in final-product analysis of mastoidectomy performance. *J Laryngol Otol.* 2015; 129(11):1091–1096. [PubMed: 26391052]
35. Andersen SAW, Foghsgaard S, Konge L, Cayé-Thomasen P, Sørensen MS. The effect of self-directed virtual reality simulation on dissection training performance in mastoidectomy. *Laryngoscope.* 2015
36. Andersen SAW, Mikkelsen PT, Konge L, Cayé-Thomasen P, Sørensen MS. Cognitive load in distributed and massed practice in virtual reality mastoidectomy simulation. *Laryngoscope.* 2015
37. Andersen SAW, Konge L, Cayé-Thomasen P, Sørensen MS. Learning Curves of Virtual Mastoidectomy in Distributed and Massed Practice. *JAMA Otolaryngol Head Neck Surg.* 2015; 141(10):913–918. [PubMed: 26334610]
38. Zhao YC, Kennedy G, Yukawa K, Pyman B, O’Leary S. Can virtual reality simulator be used as a training aid to improve cadaver temporal bone dissection? Results of a randomized blinded control trial. *Laryngoscope.* 2011; 121(4):831–837. [PubMed: 21433021]
39. Andersen SA, Cayé-Thomasen P, Sørensen MS. Mastoidectomy performance assessment of virtual simulation training using final-product analysis. *Laryngoscope.* 2015; 125(2):431–435. [PubMed: 25043448]
40. Zirkle M, Roberson DW, Leuwer R, Dubrowski A. Using a virtual reality temporal bone simulator to assess otolaryngology trainees. *Laryngoscope.* 2007; 117(2):258–263. [PubMed: 17204992]
41. Williams RG, Verhulst S, Colliver JA, Dunnington GL. Assuring the reliability of resident performance appraisals: More items or more observations? *Surgery.* 2005; 137(2):141–147. [PubMed: 15674193]
42. AAMC Careers in Medicine. Sep 20. 2015 <https://www.aamc.org/cim/specialty/list/us/336856/otolaryngology.html>

Table 1

Intra-operative mastoidectomy performance assessments

The studies listed include examples of mastoidectomy performance assessments conducted in the OR. All four reports used the Hopkins scale, and all raters were from the same institution. Studies were differentiated by those that used a Global Rating Scale (GRS) or a Task-Based Checklist (TBC). The sources of validity evidence were determined for each study based on the criteria described by Cook et al. and are listed in the final column.²¹ Each letter represents a separate evidence source as defined in Table 3.

Study	Scale	Number of Observations (Subjects)	GRS	TBC	Validity Evidence
Malik 2013 ¹⁰	Hopkins	96 (15)		x	C,D
Laceq 2012 ²⁹	Hopkins	167 (29)	x	x	D
Francis 2010 ¹	Hopkins	56 (9)		x	D,E
Francis 2010 ⁶	Hopkins	70 (15)	x	x	C,D

Table 2**Simulation-based mastoidectomy performance assessments**

The studies listed include examples of mastoidectomy performance assessments conducted using simulation. Five assessments were made in the cadaveric laboratory, ten using VR simulation, and four in both settings. The Hopkins, Iowa, OSATS, WSI, OSU, and VR assessment instruments were used. Studies were differentiated by those that used a Global Rating Scale (GRS), a Task-Based Checklist (TBC), a Final Product Analysis (FPA), or Expert Opinion (EO). The levels of validity evidence were determined for each study based on the criteria described by Cook et al. and are listed in the final column.²¹ Each letter represents a separate evidence source as defined in Table 3.

Study	Scale	Number of Observations (Subjects)	GRS	TBC	FPA	EO	Validity Evidence
Cadaveric Lab							
<i>Awad 2015</i> ²	<i>Hopkins</i>	43 (34)	x	x			<i>C,D</i>
<i>Mowry 2014</i> ¹⁷	<i>Iowa</i>	(56)			x	x	<i>D</i>
<i>Laeq 2009</i> ¹⁴	<i>Hopkins</i>	118 (23)	x	x			<i>A,C,D,E</i>
<i>Zirke 2007</i> ¹³	<i>OSATS</i>	38 (19)	x	x	x	x	<i>A,C,D,E</i>
<i>Butler 2007</i> ¹⁶	<i>WSI</i>	286 (12)			x	x	<i>A,C,D</i>
VR							
<i>Andersen 2015</i> ^{6,37}	<i>WSI</i>	920 (40)			x		<i>D</i>
<i>West 2015</i> ⁴	<i>WSI</i>	220 (22)			x		<i>D</i>
<i>Wijewickrema 2015</i> ²⁰	<i>VR/WSI</i>	48 (24)			x		<i>A,D</i>
<i>Linke 2013</i> ²⁵	<i>VR</i>	37 (37)			x	x	<i>A,D</i>
<i>Francis 2012</i> ⁵	<i>VR/OSATS</i>	96 (12)	x	x			<i>A,C,D</i>
<i>Kerwin 2012</i> ²⁴	<i>VR</i>	80 (40)			x	x	<i>A,C,D</i>
<i>Khemani 2012</i> ²³	<i>VR</i>	65 (65)					<i>A,D</i>
<i>Wan 2010</i> ¹⁰	<i>OSU</i>	N/A	x	x	x		<i>A</i>
<i>Sewell 2007</i> ⁹	<i>VR</i>	60 (15)				x	<i>A,C,D</i>
Both cadaveric and VR							
<i>Andersen 2015</i> ⁵	<i>WSI</i>	120 (40)			x		<i>D</i>
<i>Andersen 2015</i> ⁹	<i>VR/WSI</i>	68 (34)			x		<i>A,C,D</i>
<i>Zhao 2013</i> ³⁸	<i>WSI/OSATS</i>	40 (20)		x	x	x	<i>A,C,D</i>

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Study	Scale	Number of Observations (Subjects)	GRS	TBC	FPA	EO	Validity Evidence
Zirkle 2007 ⁴⁰	VR	76 (19)				x	C,D

Table 3

Validity evidence key.

Evidence Source	Letter
<i>Content</i>	<i>A</i>
<i>Response process</i>	<i>B</i>
<i>Internal structure</i>	<i>C</i>
<i>Relations with other variables</i>	<i>D</i>
<i>Consequences</i>	<i>E</i>

Table 4**Reliability**

Of the twelve studies that included reliability analysis, all but one²⁶ demonstrated inter-rater reliability with the majority utilizing raters from the same institution. Six studies also contained per-item reliability.^{14,15,16,19,24,26} Various statistical analyses were used to compute reliability.

Study	Inter-rater	Number of Raters	Single Institution?	Per-item	Statistical Analysis
Andersen 2015 ⁹	x	3*	Yes		Kappa
Awad 2015 ²	x	?	No		Concordance Correlation
Malik 2015 ¹⁰	x	4	Yes		Spearman, Intraclass Correlation
Kerwin 2012 ⁴	x	2*	Yes	x	Kappa, Spearman
Francis 2012 ¹⁵	x	5*	Yes	x	Spearman
Zhao 2011 ³⁸	x	2	Yes		Kappa, Intraclass Correlation
Francis 2010 ⁶		6	Yes	x	Cronbach α
Laeq 2009 ¹⁴	x	15	No	x	Percentage Agreement
Butler 2007 ¹⁶	x	6	No	x	Kappa
Sewell 2007 ¹⁹	x	3*	Yes	x	Pearson
Zirkle 2007 ¹³	x	2	Yes		Intraclass Correlation
Zirkle 2007 ⁴⁰	x	3*	Yes		Kappa

* These studies used both human and computer/VR-based raters.

Table 5

Frequently used acronyms.

Acronym	Definition
<i>VR</i>	<i>Virtual Reality</i>
<i>ACGME</i>	<i>Accreditation Council for Graduate Medical Education</i>
<i>ABOto</i>	<i>American Board of Otolaryngology</i>
<i>OSATS</i>	<i>Objective Structured Assessment of Technical Skill</i>
<i>GRS</i>	<i>Global Rating Scale</i>
<i>TBC</i>	<i>Task-Based Checklist</i>
<i>FPA</i>	<i>Final Product Analysis</i>
<i>EO</i>	<i>Expert Opinion</i>
<i>WSI</i>	<i>Welling Scale</i>
<i>OSU</i>	<i>The Ohio State University</i>
<i>TBD</i>	<i>Temporal Bone Dissection</i>