ORIGINAL REPORTS

# Cross-Institutional Evaluation of a Mastoidectomy Assessment Instrument

*Thomas Kerwin, PhD,* Brad Hittle, BS,* Don Stredney, MA,* Paul De Boeck, PhD,† and Gregory Wiet, MD‡,§*

*Interface Lab, Ohio Supercomputer Center, Columbus, Ohio; †Department of Psychology, Ohio State University, Columbus, Ohio; ‡Department of Otolaryngology, Ohio State University, Columbus, Ohio; and §Department of Otolaryngology, Nationwide Children's Hospital, Columbus, Ohio

**OBJECTIVE:** The objective of this work is to obtain validity evidence for an evaluation instrument used to assess the performance level of a mastoidectomy. The instrument has been previously described and had been formulated by a multi-institutional consortium.

**DESIGN:** Mastoidectomies were performed on a virtual temporal bone system and then rated by experts using a previously described 15 element task-based checklist. Based on the results, a second, similar checklist was created and a second round of rating was performed.

**SETTING:** Twelve otolaryngological surgical training programs in the United States.

**PARTICIPANTS:** In all, 65 mastoidectomy performances were evaluated coming from 37 individuals with a variety of temporal bone dissection experience, from medical students to attending physicians. Raters were attending surgeons from 12 different institutions.

**RESULTS:** Intraclass correlation scores varied greatly between items in the checklist with some being low and some being high. Percentage agreement scores were similar to previous rating instruments. There is strong evidence that a high score on the task-based checklist is necessary for a rater to consider a mastoidectomy to be performed at the level of an expert but a high score is not a sufficient condition.

**CONCLUSIONS:** Rewording of the instrument items to focus on safety does not result in increased reliability of the instrument. The strong result of the Necessary Condition Analysis suggests that going beyond simple correlation measures can give extra insight into grading results. Additionally, we suggest using a multiple point scale instead of a binary pass/fail question combined with descriptive mastery levels. (J Surg Ed ▮:▮▮▮-▮▮▮. © 2017 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

**KEY WORDS:** mastoidectomy, surgical performance evaluation, virtual reality simulation, assessment

**COMPETENCIES:** Medical Knowledge, Practice-Based Learning and Improvement

## INTRODUCTION

For the results of performance tests to be valuable for making decisions, care must be given to understand the characteristics of that test. Using a poorly understood and unreliable performance test in a decision-making process can be worse than no test at all, since it gives unfounded confidence in that decision. Before adoption into a curriculum or use in certification, a surgical performance instrument must be thoroughly reviewed. Validity frameworks by Messick[1] and Kane[2] give a structure for evaluating the validity of measurement instruments in a rigorous way. Both of these frameworks emphasize the importance of a coherent argument toward the use of a measurement instrument for a particular purpose. The objective of this work is to obtain validity evidence for an universal evaluation instrument used to assess the performance level of a mastoidectomy. We believe the procedures described here are easily adapted to other surgical performance instruments, although the work involved in creation and evaluation of a particular instrument will always be substantial.

Many surgical performance instruments are developed and tested at a single institution or in a small geographical area. The 2 instruments examined in the current study were developed with input from experts in mastoidectomy from

This work was supported by The National Institute for Deafness and other Communication Disorders, National Institutes of Health, USA, R01DC011321.

*Correspondence*: Inquiries to Thomas Kerwin, PhD, 1224 Kinnear Road, Columbus, OH 43212; e-mail: kerwin@osc.edu

  1931-7204/$30.00 http://dx.doi.org/10.1016/j.jsurg.2017.10.006 

multiple institutions across the United States. Gathering consensus on the important qualities of a successful surgical procedure from a wide range of experts helps to minimize the personal differences in technique and didactic focus that could be concentrated at a single institution. A study by Wan et al.[3] developed a set of "universal metrics" based on a literature review and then rank them in terms of importance through a survey of 2 national otology societies. Using our expert consortium (14 individuals from 12 different institutions), the individual items from the Wan et al. study were explicitly defined so that a uniform interpretation could be applied for determining success or failure (binary decision process) for each item. Using a virtual reality temporal bone simulator system, we administered the original instrument and also altered it to make a second instrument that focused on safety.[4] We acquired a wide selection of mastoidectomies by experts, residents, and medical students (MS) from 12 institutions, and those mastoidectomies were evaluated using the 2 instruments.

We examine reliability measures and discuss validity evidence in using the 2 instruments tested to make judgments about skill levels. Necessary condition analysis (NCA) is introduced as an appropriate technique for evaluating relationships between performance variables that may not be captured well by existing methods.[5] Additionally, we compare our results with studies examining other mastoidectomy evaluation scales, especially a scale developed at John Hopkins by Francis et al.[6] Work by Sethia et al.[7] provides an overview of this and other instruments and points out that existing instruments have been developed and tested at only a small number of institutions. Our scale is similar to the Hopkins scale: both have a task-based checklist (TBC) and a global rating scale (GRS). Five of the 22 items in the Hopkins scale TBC are nearly the same as in ours, but the phrasing and the content of the other items differ. Also, our scale has only a single question GRS where they have 10 items. A major difference between this work and others, including those using the John Hopkins instrument, is that the number of institutions involved in both the development and application of the instrument is much larger in our work. Finally, our findings will be discussed in terms of reliability and validity, using Messick's[1] framework for the latter.

## MATERIALS AND METHODS

This study was approved (ID 2011H0253) by both The Ohio State University Office of Responsible Research biomedical institutional review board (IRB) as well as by the IRBs of each local institution involved in the study. A click-through consent form was part of the software.

### Simulation and Grading Environment

The surgical simulation system that was used to gather the mastoidectomy and adapted to provide a grading environment for the virtual mastoid surgeries is discussed in Wiet et al.[8] The system presents a virtual temporal bone in three-dimensional space. The temporal bone data was acquired using micro-CT and 3 different virtual bones were used in this study. All 3 appeared healthy (i.e., nonpathological). The bones are viewed by the users with active 3D glasses to provide a stereoscopic image of the bone as one would see through the operating microscope. Two haptic joysticks (with 6 degree-of-freedom movement) are used to control the drill and suction-irrigation device. Users may manipulate bone orientation, change magnification, and select different drill burr sizes and types when they are performing the virtual surgery. Performances are recorded for playback and review. Grading was performed on the same hardware using a program that could play back the mastoidectomy performances for the expert reviewers. The software includes the ability to view sections of the procedure multiple times and also to pause the playback and rotate the virtual bone, viewing it from different angles. The reviewers selected pass or fail on the list of items to the side of the bone display. Based on a previous request from reviewers, to decrease the time needed for grading, the virtual dissection was played back at double speed.

### Study Execution

Twelve sites had been previously equipped with our simulator system. The participating sites all have ACGME accredited residency education programs in otolaryngology. Residents and faculty at all sites used the simulation environment to perform 3 complete mastoidectomies including facial recess dissection. The 3 surgeries were performed on separate virtual bones, but each participant had the same set of 3 bones. The participants cover a wide distribution of skill levels: medical students (MS), postgraduate year (PGY) 2-5, fellows, and attending physicians (experts).

In all, 249 data files were created by the participants, 83 of those were adequate for analysis: the others were false starts or incomplete data. In the simulation, a series of steps for the mastoidectomy were indicated and the users went through them, pressing "next" each time. If all the steps were not indicated by the user as being completed, the dataset was ignored. Also, datasets where no drilling was performed were ignored.

Out of those 83, 66 were selected randomly to give an even distribution over experience levels and to give each of the 12 reviewers 11 mastoidectomies to review. The burden for review was high, since it could take up to 30 minutes in some cases to review 1 mastoidectomy. In this distribution, not all 3 mastoidectomies performed on the different bones from each participant was selected: 23 participants had 1 performance selected, 2 participants had 2 performances selected, and 13 had 3 performances selected.