PeerJ Preprints NOT PEER-REVIEWED

Cross-institutional evaluation of a mastoidectomy assessment instrument

Thomas Kerwin, PHD[1], Brad Hittle, BS[1], Don Stredney, MA[1], Paul De Boeck, PHD[2], Gregory Wiet, MD[3,4]

[1] Interface Lab, Ohio Supercomputer Center, Columbus, Ohio, United States

[2] Department of Psychology, Ohio State University, Columbus, Ohio, United States

[3] Department of Otolaryngology, Ohio State University, Columbus, Ohio, United States

[4] Nationwide Children's Hospital, Columbus, Ohio, United States

Corresponding Author:

Thomas Kerwin[1]

1224 Kinnear Rd, Columbus, Ohio, 43212, United States

Email address: kerwin@osc.edu

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

PeerJ Preprints NOT PEER-REVIEWED

Cross-institutional evaluation of a mastoidectomy assessment instrument

Abstract

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

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 (ICC) 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.

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 towards 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 two instruments examined in the current study were developed with input from experts in mastoidectomy from 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 two 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 from 12 institutions, and those mastoidectomies were evaluated using the two instruments.

We examine reliability measures and discuss validity evidence in using the two instruments tested to make judgments about skill levels. Necessary condition analysis 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 ten 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 microCT and three different virtual bones were used in this study. All three appeared healthy (i.e. non-pathological). 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 three complete mastoidectomies including facial recess dissection. The three surgeries were performed on separate virtual bones, but each participant had the same set of three bones. The participants cover a wide distribution of skill levels: medical students, Post-Graduate Year (PGY) 2-5, fellows and attending physicians (experts).

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 twelve reviewers eleven mastoidectomies to review. The burden for review was high, since it could take up to 30 minutes in some cases to review one mastoidectomy. In this distribution, not all three 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 three performances selected.

Each of twelve expert reviewers, all considered experts in otologic surgery, was assigned eleven grading tasks (individual mastoidectomy performances). They were blinded to the identity of the subject performing the dissection and did not review their own performances. This resulted in two gradings for each virtual mastoidectomy in the testing set. In the first trial, one expert failed to evaluate a particular performance (by a PGY5 who had only one performance selected to be graded), so we eliminated that performance from the data, leaving 65 performances, with a total of 130 evaluations for each trial. The total set evaluated in the current study comprises 38 sessions collected from faculty and 27 collected from fellows, residents and medical students (MS) (Expert = 38, Fellow = 3, MS = 1, PGY1 = 3, PGY2 = 4, PGY3 = 5, PGY4 = 5, PGY5 = 3, PGY6 = 3).

The reviewers were also asked to give a subjective assessment of the level of training that the mastoidectomy performance represents (a type of global rating). For the subjective assessment, the global rating choices were: novice, intermediate, and expert. Novice level was defined as "ready for the temporal bone lab", Intermediate level as "ready for real patients in the operating room" (under supervision) and Expert level as "ready to operate without supervision."

Rating instrument

We used the two rating instruments described in Kerwin et al.[4] The instruments cover the technical skills used in two-handed surgical tool manipulation and bone removal in a mastoidectomy. Ratings did not happen at the same time: all ratings from the first instrument were collected several months before the ratings from the second instrument. The second instrument is a revised version of the first that emphasizes safety in the phrasing of the items.

As noted above, two trials were completed, with different performances assigned to the experts and different sets of evaluation items. The two evaluation sets of items are related but not identical. The first trial had a list of 16 items adapted from the work of Wan et al.[3] In the second trial, based on feedback from the expert reviewers and an additional Delphi method, we attempted to more sharply define the assessment items in more universal terms, emphasizing safety. Additionally, at the suggestion of the expert group, two of the items were combined into one. This means that the second trial used a list of 15 items; item number 10 was removed from the list but the numbers of the other items remained the same. The text of the items for both trials can be seen in Table 1. All item specific ratings were binary in terms of pass (=1) and fail (=0). A total instrument score was calculated by counting the number of items given a pass rating.

To ascertain the evidence for validity of the instrument, we use inter-rater reliability measures, correlations between scores and experience and necessary condition analysis, all which are described with the results of those techniques in the next section.

Table 1: Text of questions asked during mastoidectomy performance review. Question #10 in the first trial had no corresponding question in the second trial.

Number	Trial 1	Trial 2
1	Maintains visibility of burr while removing bone	Maintains safe view of the burr throughout the procedure
2	Excessive force will not be used near critical structures	Maintains safe force near critical structures throughout the procedure
3	Appropriate depth of cavity	Sufficient removal of mastoid air cells for proper visualization of deep structures

4	No holes in tegmen	Maintains integrity of tegmen
5	Select appropriate burr	Efficient and Safe burr selection
6	Violation of the sigmoid sinus	Maintains integrity of sigmoid sinus
7	Identification of chorda tympani nerve	Identifies chorda tympani nerve sufficiently to perform facial recess approach
8	Drill in best direction	Efficient and safe direction of drilling (parallel to critical structures)
9	External auditory canal wall will remain up	Sufficient thinning of posterior external auditory canal wall to visualize facial nerve
10	No holes in external auditory canal wall	
11	Complete saucerization	Sufficient saucerization for safe drilling
12	Posterior external auditory canal wall thinned appropriately	Avoids overthinning or holes in posterior auditory canal wall
13	Violation of the facial nerve	Maintains integrity of facial nerve
14	Violation of the horizontal (lateral) semi-circular canal	Maintains integrity of horizontal semi-circular canal
15	Drill contact with ossicles	Maintains integrity of ossicles
16	Violation of dura	Maintains integrity of dura

Results

Inter-rater reliability

Several measures of inter-rater reliability were calculated, both per-item and using the total checklist score. Percentage agreement, intraclass correlation (ICC)[9], and Cohen's kappa per-item for both trials are shown in Table 2 and Table 3 and discussed below. Pass percentages are included since very high or low numbers of passing grades can lower the utility of inter-rater reliability statistical measures. The ICCs are also presented in Figure 1, where the confidence intervals are shown.

Table 2: Per-item inter-rater reliability statistics and pass percentages from Trial 1.

Question Number	Pass Percentage	ICC(2,1)	Cohen's Kappa	Percentage Agreement
1	0.66	0.33	0.32	69.2

PeerJ Preprints NOT PEER-REVIEWED

2	0.66	0.49	0.46	75.4
3	0.67	0.27	0.27	67.7
4	0.55	0.54	0.54	76.9
5	0.46	0.01	0.01	47.7
6	0.64	0.57	0.57	80.0
7	0.51	0.18	0.17	56.9
8	0.52	0.10	0.10	53.9
9	0.82	0.32	0.30	78.5
10	0.72	0.45	0.44	76.9
11	0.54	0.07	0.07	53.9
12	0.55	0.20	0.19	60.0
13	0.67	0.48	0.48	76.9
14	0.84	0.62	0.61	89.2
15	0.84	0.15	0.15	76.9
16	0.66	0.26	0.25	66.2

Table 3: Per-item inter-rater reliability statistics and pass percentages from Trial 2.

Question Number	Pass Percentage	ICC(2,1)	Cohen's Kappa	Percentage Agreement
1	0.60	0.18	0.16	56.9
2	0.58	0.14	0.13	56.9
3	0.54	0.01	0.01	50.8
4	0.55	0.46	0.45	72.3
5	0.46	0.13	0.13	56.9
6	0.58	0.57	0.56	78.5
7	0.38	0.18	0.18	61.5
8	0.47	0.09	0.08	52.3
9	0.71	0.12	0.12	63.1
10	NA	NA	NA	NA
11	0.63	0.08	0.08	56.9
12	0.69	0.57	0.57	81.5
13	0.66	0.52	0.52	78.5

PeerJ Preprints NOT PEER-REVIEWED

14	0.85	0.35	0.34	83.1
15	0.82	0.11	0.11	73.8
16	0.69	0.43	0.42	75.4

The intraclass correlation (ICC) is a common reliability measure that compares the variance from consistency between raters with the total variance. For cases with no agreement, the ICC would be 0 and in cases where there is total agreement, the ICC would be 1. In this work, each surgical example is rated by two experts. There is some overlap between the raters; each rater in our group did not rate all bones, but rated a set of them. This falls under case 2 in Shrout and Fleiss's definition of the ICC[9], but with incomplete data. We use the ICC(2,1) formulation of the measure. As seen in Figure 1, the ICC results vary greatly between 0 and 0.6 for each item and some are very low. For the subjective (global rating) item, since the question was identical in both trials, we can compute the reliability ($ICC(2,1) = 0.39$) for that rating across four raters, instead of two. (not shown in Table 2 or Table 3 but in Figure 1 instead)

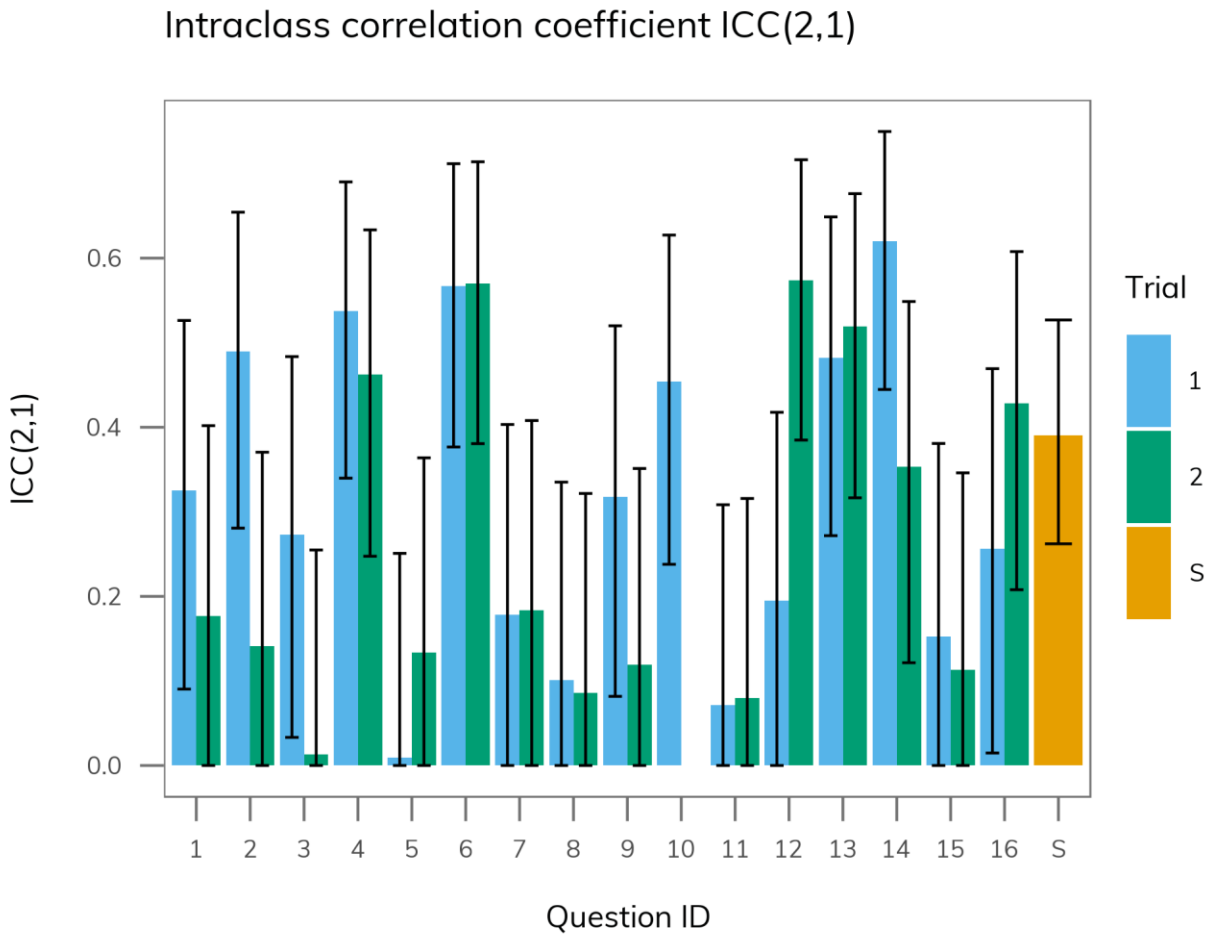

Figure 1: Comparison of intraclass correlation ICC(2,1) for all evaluation items. 95% confidence intervals are marked. The value of ICC(2,1) for subjective assessment across both trials is shown as index S. Trial 2 did not have a question 10.

As mentioned above, we determine a total instrument score by tallying the positive responses. ICC(2,k) for the total instrument score is 0.59 for trial 1 and 0.46 for trial 2.

Correlations and NCA

Since the individual year groups each contain a small number of participants, we consider three experience levels instead: a medical student through a PGY3 has low experience, a PGY4 through a fellow has moderate experience, and a faculty member has high experience. Spearman's rho is used to judge the strength of the relationship between instrument score, experience level, and global rating. The correlation between the total instrument score and the global rating is strong: Trial 1, $\rho_s = 0.66, p < 0.01$; Trial 2, $\rho_s = 0.75, p < 0.01$. The correlation between the experience level of the individual and the global rating is low and not significant for the second trial: Trial 1, $\rho_s = 0.21, p = 0.017$; Trial 2, $\rho_s = 0.17, p = 0.055$. The correlation between the experience level of the individual and the total instrument rating is low: Trial 1, $\rho_s = 0.27, p < 0.01$; Trial 2, $\rho_s = 0.21, p = 0.019$. Box-plots comparing the distribution of the total instrument rating for the three experience levels can be seen in Figure 2.

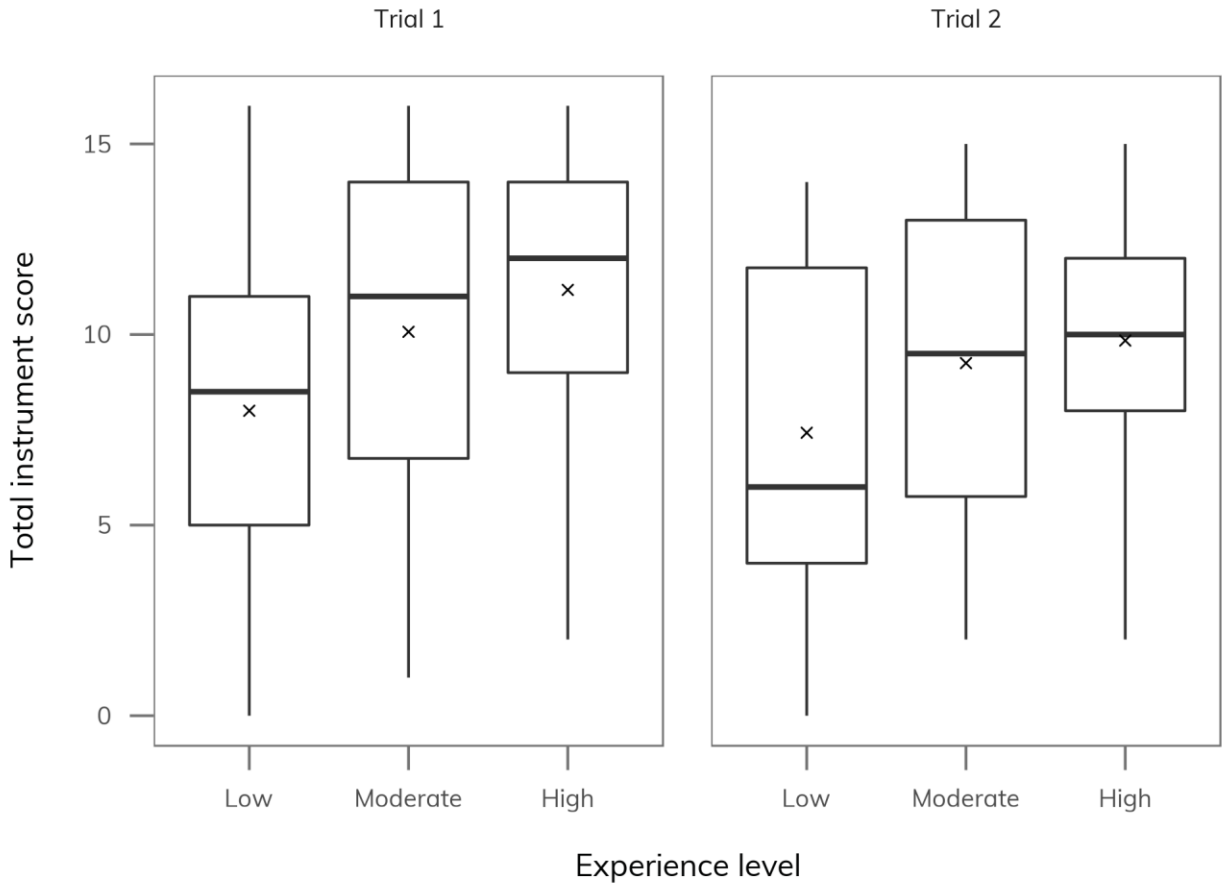

Figure 2: Boxplots showing the distribution of total instrument scores for the three experience levels. The middle line of the boxplot shows the median value while the mean is designated by an 'X'.

Figure 3 shows the distributions of the total instrument score for different values of the global rating score. Examining the graph, high scores were given to mastoidectomies considered "expert" level, moderate to high scores were given for those considered "intermediate" level and the full range of scores were given to those considered "novice" level. Based on this observation, we can say that a high total instrument score is a necessary but not sufficient condition for considering a mastoidectomy performance as one of higher skill, as determined by the global rating. Additionally, a low score is a sufficient condition to be considered a novice.

A necessary condition is one that needs to be present for a specified outcome to come into effect. In our case, a mastoidectomy needs a high total instrument score for that mastoidectomy to be considered one of high skill by the raters. Necessary condition analysis[5] (NCA) is a recent technique that assists in evaluating claims about this type of necessary condition relationship. NCA uses a "ceiling line" to define the amount of empty space in the upper left of a scatterplot and uses this to calculate an effect size. In NCA, the size of the range of the data is given by one

number, *scope*, and the amount of empty space on the upper left is calculated as the *ceiling zone*, and then a ratio is found to determine the strength of the relationship. For Trial 1, NCA yields 0.5 and 0.57 for Trial 2. Effect sizes between 0.3 and 0.5 are considered medium and those above 0.5 are considered large. An NCA test was also performed examining a potential relationship between the total instrument score and the PGY level of the person who performed the mastoidectomy (including levels for medical students and attendings). The tests for those did not show a strong effect: 0.094 for Trial 1 and 0.13 for Trial 2. Table 4 contains further details from the NCA.

Table 4: Results of Necessary Condition Analysis for the sum of item checklist scores as the independent variable and the global rating score and PGY level as the dependent variables. Effect sizes above .5 are considered strong.

	Trial 1 (Global Score)	Trial 2 (Global Score)	Trial 1 (Exp. Level)	Trial 2 (Exp. Level)
Accuracy	100.0	100.00	100.00	100.00
Ceiling zone	16.0	17.00	3.00	4.00
Effect size	0.5	0.57	0.09	0.13
Scope	32.0	30.00	32.00	30.00

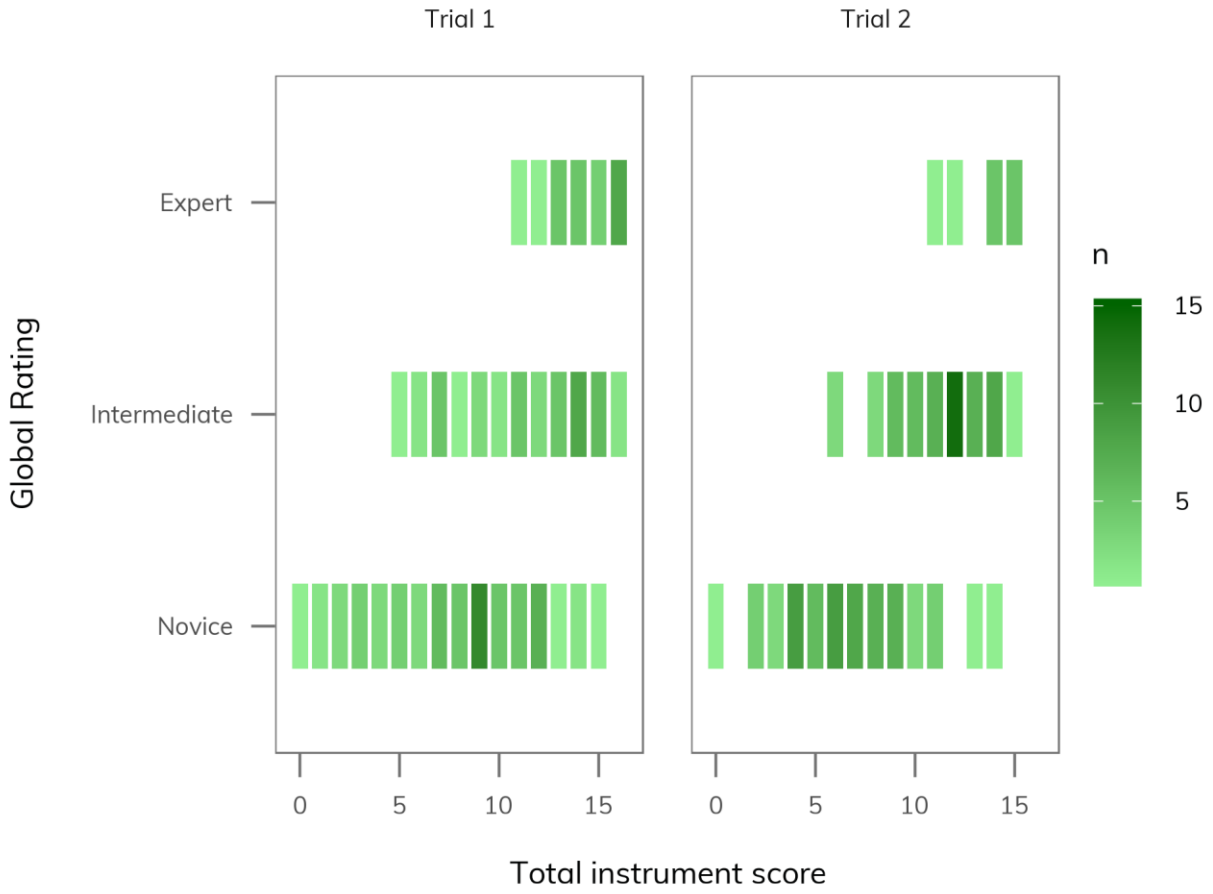

Figure 3: Distribution of the sum of item checklist scores among different global rating categories. Each rectangle represents a particular result score for that global rating level. Darker rectangles indicate more grades falling into that bin. Higher subjective scores tend to receive higher item checklist scores.

Discussion

Data acquisition

The population of medical residents training in otolaryingolgy is not a large one, and that fact can make data acquisition from that group difficult. Our study took place in many institutions but the participants were a sample of convenience from those institutions. This might lead to unknown bias. Additionally, the fact that many of the mastoidectomies rated were performed by the same individual on different bones could introduce some bias into the scores for each experience level, since two mastoidectomies performed by the same person can be assumed to have related scores. This is a limitation of how the data were collected for this study.

Comparison with other studies

A number of studies have investigated reliability and validity evidence of mastoidectomy evaluation instruments. We discuss the statistical measures used to support reliability and validity claims about the more prominent instruments. Many are based on the "Objective structured assessment of technical skill" (OSATS) framework introduced by Martin et al.[10]. The OSATS framework is very popular and can be a useful tool in developing an instrument. However, the mere fact of using the OSATS framework to develop an instrument does not mean that that instrument is valid or reliable. Reliability and validity evidence must be evaluated separately for each test instrument.

Using the OSATS framework, Johns Hopkins researchers developed an instrument for mastoidectomy performance, containing both a Task-Based Checklist (TBC) and Global Rating Scale (GRS)[6]. Assessment was conducted by expert evaluators watching resident performances in the OR. Raters were not blinded to the subject's identity. They found moderate correlations between days in the otology program and both the overall TBC score ($r = 0.60$) and the overall GRS score ($r = 0.57$). Correlation between GRS and TBC was very high ($r = 0.93$). No mastoidectomies from attending surgeons were included and inter-rater reliability was not measured.

Laeeq et al.[11] conducted a test of the Johns Hopkins scale evaluating resident performance in the temporal bone lab. By assigning a value of "pass" to items scored 3, 4, and 5 and "fail" to items scored 1 or 2, they showed pass/fail percentage agreement values per item on their TBC ranging from 54% to 86%, with most items in the 70% to 80% range. They did not report correlation between the TBC and GRS, but correlation between the TBC and one item ("Overall surgical performance") was moderate ($r = 0.69$). They did not provide kappa or ICC scores. No mastoidectomies from attending surgeons were included. Performance on their instrument significantly increased based on level of experience as determined by ANOVA, but there is no report on the strength of the association. Experts were not blinded to the identity of the resident.

More recently, Awad et al.[12] evaluated the use of the Hopkins instrument for resident performance in the temporal bone lab. They are notable as the first "outside" group to use the Hopkins instrument. They showed a significant positive correlation with training level for both the TBC and GRS using the Spearman rank correlation coefficient; weakly for the TBC ($r_s = 0.117$) and moderately for the GRS ($r_s = 0.330$). They used between two and four raters per evaluation and report that the "interassessor concordance was high, ranging from 70% to 80%". From the context, this seems to be referring to percentage agreement per item but it is not

entirely clear. Experts were not blinded systematically to the identity of the resident, but it is not clear if they knew the experience level of the resident prior to grading.

As seen in the previous three examples, the Hopkins scale showed impressive results when administered by experts at the same institution it was developed, but subsequent applications by other experts showed considerably more modest results. Our current work differs in various important respects: our raters are blinded systematically to the identity of the person who performed the mastoidectomy, our application of the instrument is on participants with a wide range of experience levels, and our raters are from a large group of institutions. Our experimental design leads to results that are more generalizable than earlier work.

As mentioned earlier, we do not advocate the use of percentage agreement as a measure of inter-rater reliability. However, we do present it with the ICC for comparison to other studies. Our results for individual items are similar to those in the reports of Laeeq et al.[11] and Awad et al.[12]

Other instruments, such as the Welling Scale and the one by Zirkle et al.[13] have also been developed, and a review can be read in Sethia et al.[7]. The checklists in the current work are similar to those of the Johns Hopkins assessment. Generally speaking, the individual items used in the available methods of mastoidectomy assessment reported in the literature have not been shown to have excellent reliability or validity.[7] Assessment instruments can be used for summative and for formative purposes. For summative performance the total score is used but the current tools seem far from providing enough evidence for high stakes judgments to be routinely made based on the results. For formative feedback, assessments are important during training and necessary for adequate technical skill development.[14] Formative feedback depends on communicating to the trainee both what is being done correctly and incorrectly. For an instrument to be effective in this application, individual items must each show both reliability and validity to the construct of mastoidectomy surgery. With valid and reliable individual items, performance on specific items becomes the basis for this feedback.

For our current instruments to be universally accepted for both summative and formative applications, a significant uphill road lies ahead to provide sufficient reliability and validity evidence. Use of more modern psychometric techniques such as Item Response Theory may provide the framework to achieve this level of evidence[15]. For a testing instrument to be feasible to implement, we must be able to use any small group of skilled raters to administer the instrument. This can be a high bar and this type of evidence can be difficult to obtain.

Reliabilty

Reliability is a prerequisite for validity and can be examined in numerous ways. As mentioned above, earlier studies use percent agreement scores to gauge inter-rater reliability, but these can be misleading, especially in the case of test items that have very high or very low pass rates. We encourage the use of ICC for this measurement, since it a flexible measure.

The questions associated with low ICC values in both of the two trials are ones associated with burr selection and drilling direction, identification of the chorda tympani nerve, saucerization and drill contact with ossicles. Questions that had high ICC values in both trials include violation of the sigmoid sinus and violation of the facial nerve. It is not surprising that these two violation questions have high inter-rater reliability, since they are common errors in learning the procedure (resulting in graders looking out for those errors specifically) and are obvious when they occur. The software used by graders to look at the drilled bone highlighted regions of critical structures (sigmoid sinus, facial nerve, dura and lateral semi-circular canal) that were removed in the course of the procedure. The fact that this automatic highlighting gave a visual representation of the amount of violation that occurred probably contributed to the high inter-rater reliability of these items.

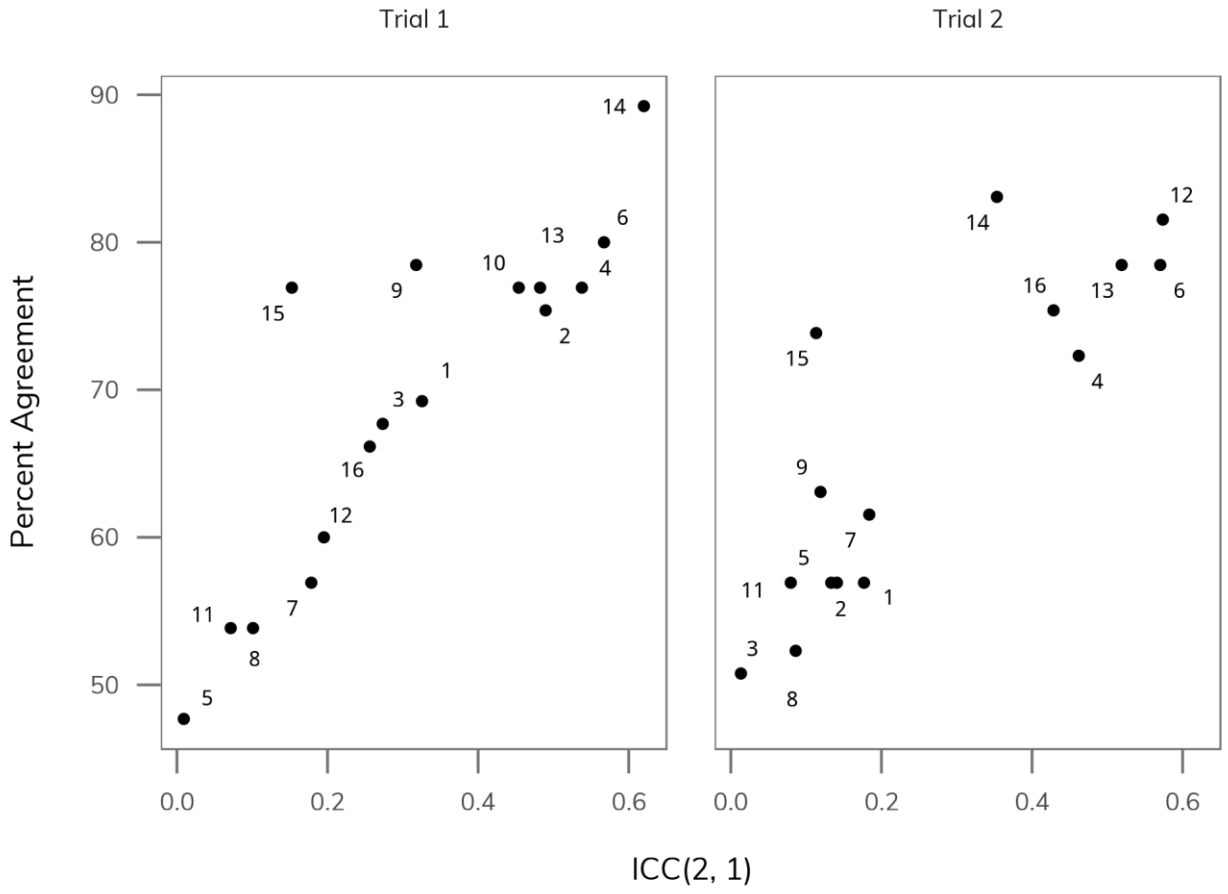

Figure 4: Comparison of Percentage agreement and ICC values for each item for both trials.

Figure 4 displays the relationship between percentage agreement and the ICC values for both trials. Question 15 has a much higher percent agreement score than ICC, relative to the other items. Question 15 concerns drilling on ossicles, which was a rare occurrence. Percent agreement is high because the majority of the answers were true, indicating that no ossicle was hit. Because only occasionally a performance was a failure, the item does not yield enough information to evaluate inter-rater reliability. These types of situations show how percentage agreement is not suitable to be used alone for inter-rater reliability evaluation.

In our first experiment, we found that our initial application of the instrument demonstrated only moderate reliability. In reviewing the results with the expert graders, a consensus was reached that the disagreement between expert raters was perhaps due to differing preferences in technique rather than emphasizing safe surgical technique. We therefore performed an additional Delphi process in which the definitions of each item were further refined based on the ultimate premise that they would be used to identify "safe" as opposed to "proper" surgical technique. The rationale was that experts would more easily agree on what surgical technique was considered safe rather than what was the best technique possible. This modified instrument was used in the second experiment. However, the second instrument was not more reliable than the first.

Using binary pass/fail scores are perhaps not optimal compared with a rating scale with multiple points on the scale. Not only is there an aspect of subjective decision making in all of the questions, each grader deciding their own threshold between pass and fail slightly differently, but there is also less information extracted from the raters. Asking raters to respond on a larger scale range, we could obtain more information about individual graders and perhaps factor their individual biases into account for a final grade. Using more than two raters could also increase reliability in the scores. However, the number of potential raters is small, since they must be well experienced in mastoidectomy technique. This makes averaging over a larger group of raters infeasible in practice.

Validity evidence

When talking about validity of a particular test, Kane pointed out that it is a two step process: consider the specific purpose a test will be used for and then develop the argument that the test will be useful for that purpose.[2] Although a further goal is high-stakes assessment, what we propose here is the use of the checklists for feedback while residents are learning techniques in a temporal bone lab, virtual or otherwise. Many other investigations of instrument validity use the elements of Messick's framework of validity[1] to categorize elements of evidence for validity.

Messick identified six aspects of construct validity and we believe the development and testing of our assessment instrument touches on all of them.

The content aspect concerns the fact that the assessment would cover all parts of the domain in question. The development of the test[4] involved experts considering all aspects of the technical skills used in mastoidectomy and paring down the list to the ones they collectively considered the most important.

The substantive aspect involves incorporating tasks in the assessment that sample the real life thing that is to be measured. In our case, we use a computer simulation of a mastoidectomy, but the tasks that are to be performed in the assessment are well handled by the simulation. Furthermore, it is a simulation of a surgery as opposed to a simple box trainer or an isolated specific sub-task. The correlation between the total instrument score and the global rating score provides further evidence that the overall opinion of the experts matches with the tasks performed in the assessment.

Generalizability concerns the tasks and populations to which the assessment is applicable. The participants performed the mastoidectomy on one of three different virtual mastoid bones, all from healthy adults. Mastoidectomy on pathological bones was not tested. Participants from a wide range of skill levels and from many institutions, lending evidence to this aspect.

We use Necessary Condition Analysis to provide evidence for validity through the structural aspect. For this aspect, the relationship between the instrument score and the construct is investigated. The results of the NCA show a strong effect suggesting there are necessary conditions to be considered an expert and sufficient conditions to be considered a novice. Both checklists are capturing aspects of mastoidectomy skill that are necessary to be considered an expert. No mastoidectomies that were considered expert level got a low score on the checklist. However, individual performances that are considered "novice" can still get high scores on the checklist, while a low score is sufficient to be considered a novice. An interpretation of this result is that while novices might perform well, experts almost always perform well. The global rating scale of the performance could be influenced by tool motion that appears, subjectively, more trained. A careful novice may succeed in individual tasks but still *look* like a novice, but a more skilled individual would both look and act in a skilled manner.

The external aspect of validity can be tested by looking at the relationship between the instrument results and other measures of the subjects. Our evidence for a relationship with assumed surgical training based on experience level for this instrument is low in spite of the strong evidence of a relationship between the global score and item checklist. The global score and the checklist score were more consistent with each other than with the participants' experience level. This may be due to lack of high stakes testing (i.e. performing the

mastoidectomy may not have been taken seriously enough), lack of proficiency in using the simulator system, or lack of fidelity in some parts of the simulator or could perhaps show that the relationship between years in training and performance is not as strong as traditionally accepted.

The consequential aspect of validity concerns the effect of the use of the instrument. Since there was no feedback loop present where experts or trainees could view the scored instrument or be affected by it in anyway, there is no evidence one way or the other for this aspect. However, the results from the NCA imply that this instrument should be used as a low-bar "screener", rather than using the total instrument score to precisely judge competence.

Conclusion

To our knowledge, this study reports the results of the first attempt to test a rating instrument for mastoidectomy skill across more than 10 institutions. Drilling performances were obtained from a wide range of skill levels from the 12 different institutions. The instruments showed very strong evidence for a necessary condition relationship but low to moderate ICC values. Reliability measures were not higher for the instrument that focused on safety. Achieving high inter-rater reliability could be more difficult with raters at many institutions due to differences in didactic focus and technique between those institutions. We feel that use of NCA can be used as an companion technique to traditional correlation analysis to examine the validity of screening instruments to establish a minimum skill level.

Our ultimate goal is to have a scoring instrument for mastoidectomy that is useful in high-stakes assessment (e.g. board certification). A limitation of our study is that the validity evidence found is not strong enough to support that use. The adjustments to the item texts that were made to emphasize safety did not significantly change reliability measures. Additionally, the uneven distribution of mastoidectomy performances from different skill levels is a sub-optimal feature of our study to reveal differences between skill levels. However, we have shown that there can be great difficulties in developing scoring instruments that can be used with multiple raters, multiple experience levels and multiple institutions. Assessment tools that directly affect the career of surgeons need to be tested in real-world conditions and challenged before use for decision making.

We plan to improve the instrument using the information obtained from this study and after further refinement and vetting, we hope that such an instrument will have great utility for use in cross-institution curricula and certification for otologic surgery. Additionally, the process described here can be honed and adapted to gather validity evidence for any instrument designed for the evaluation of surgical skills, keeping in mind differences between raters and institutions.

Acknowledgements

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

Bibliography

1. Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*. 1995;50(9):741-749. doi:10.1037/0003-066X.50.9.741.

2. Kane M. The argument-based approach to validation. *School Psychology Review*. 2013;42(4):448+.

3. Wan D, Wiet GJ, Welling DB, Kerwin T, Stredney D. Creating a cross-institutional grading scale for temporal bone dissection. *Laryngoscope*. 2010;120:1422-1427. doi:10.1002/lary.20957.

4. Kerwin T, Hittle B, Stredney D, De Boeck P, Wiet G. Multi-institutional development of a mastoidectomy performance evaluation instrument. *Journal of Surgical Education*. May 2017. doi:10.1016/j.jsurg.2017.05.006.

5. Dul J. Necessary condition analysis (NCA): Logic and methodology of "necessary but not sufficient" causality. *Organizational Research Methods*. 2016;19(1):10-52. doi:10.1177/1094428115584005.

6. Francis HW, Masood H, Chaudhry KN, et al. Objective assessment of mastoidectomy skills in the operating room: *Otology & Neurotology*. 2010;31(5):759-765. doi:10.1097/MAO.0b013e3181e3d385.

7. Sethia R, Kerwin T, Wiet GJ. Performance assessment for mastoidectomy: State of the art review. *Otolaryngology-Head and Neck Surgery*. 2017;156(1):61-69. doi:10.1177/0194599816670886.

8. Wiet GJ, Stredney D, Kerwin T, et al. Virtual temporal bone dissection system: OSU virtual temporal bone system: Development and testing. *Laryngoscope*. 2012;122 Suppl 1:S1-12. doi:10.1002/lary.22499.

9. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-428.

10. 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.

11. Laeeq K, Bhatti NI, Carey JP, et al. Pilot testing of an assessment tool for competency in mastoidectomy. *Laryngoscope*. 2009;119:2402-2410. doi:10.1002/lary.20678.

12. Awad Z, Tornari C, Ahmed S, Tolley NS. Construct validity of cadaveric temporal bones for training and assessment in mastoidectomy: Validity of CTB for mastoidectomy training. *The Laryngoscope*. 2015;125(10):2376-2381. doi:10.1002/lary.25310.

13. Zirkle M, Taplin MA, Anthony R, Dubrowski A. Objective assessment of temporal bone drilling skills. *Annals of Otology, Rhinology & Laryngology*. 2007;116(11):793-798. doi:10.1177/000348940711601101.

14. Ericsson KA. Deliberate practice and acquisition of expert performance: A general overview. *Academic Emergency Medicine*. 2008;15(11):988-994. doi:10.1111/j.1553-2712.2008.00227.x.

15. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education: Classical test theory and item response theory. *Medical Education*. 2010;44(1):109-117. doi:10.1111/j.1365-2923.2009.03425.x.