Published in final edited form as:

Eval Health Prof. 2008 December ; 31(4): 419–436. doi:10.1177/0163278708324444.

Reliability of Surgical Skills Scores in Otolaryngology Residents Analysis Using Generalizability Theory

Soledad A. Fernandez
Ohio State University, Columbus

Gregory J. Wiet
Columbus Children's Hospital, Columbus, Ohio Ohio State University

Nancy N. Butler, **Bradley Welling**, and **David Jarjoura**
Ohio State University, Columbus

Abstract

Assessments of temporal bone dissection performance among otolaryngology residents have not been adequately developed. At the Ohio State College of Medicine, an instrument (Welling Scale, Version 1 [WS1]) is used to evaluate residents' end-product performance after drilling a temporal bone. In this study, the authors evaluate the components that contribute to measurement error using this scale. Generalizability theory was used to reveal components of measurement error that allow for better understanding of test results. A major component of measurement error came from inconsistency in performance across the two cadaveric test bones each resident was assigned. In contrast, ratings of performance using the WS1 were highly consistent across raters and rating sessions within raters. The largest source of measurement error was caused by residents'inconsistent performance across bones. Rater disagreement introduced only small error into scores. The WS1 provides small measurement error, with two raters and two bones for each participant.

Keywords

measurement error variance; *mixed models*; *performance measures*

Human temporal bone dissection for the otolaryngology resident is a core component of surgical skills training. The temporal bone encompasses the most lateral aspect of the skull and includes the area of the jaw joint and the external ear, extending back to the occipital bone and the area just beneath the ear, known as the lateral skull base, that continues to the cervical spine. This region measures about 5 cm (Shah & Darzi, 2001). Unlike most surgical areas, the temporal bone houses a high density of vital structures. These include the organs for the sense of hearing and balance, the nerves for facial movement, and the major blood vessels to and from the brain. The superior boundary of the temporal bone is the temporal lobe of the brain; the medial boundary is the brainstem and cerebellum. Temporal bone dissection is a core surgical skill in ear, nose, and throat surgery. It is required to treat various diseases of the ear and brain such as infection, tumor removal, congenital disorders, deafness, and balance disorders, to name a few. As with most surgical techniques, temporal bone dissection requires a consummate understanding of the anatomical contents of the area

Address correspondence to Soledad Fernandez, 320 West 10th Ave., M410 Starling Loving Hall, Columbus, OH 43210; Soledad.Fernandez@osumc.edu..

as well as technical skill manipulating instruments and tissues. The high density of vital structures requires surgical techniques with sub-millimeter accuracy. This necessitates development of microsurgical techniques: that is, skill operating under a microscope. Compounding this entire scenario is the fact that all these structures are embedded within bone and surgical access is achieved by systematic removal of bone around each vital structure. Use of different-sized and -consistencies drill bits under a steady stream of fluid to remove debris is required when dissecting. Surgical misadventure in this area can result in deafness, imbalance, facial paralysis, injury to the brain, and death.

Given this background, it is evident that temporal bone dissection training is not something that is mastered easily. In reality, it requires several years of training to become proficient and many years and numerous surgical cases to develop to the expert level. Ear surgery is one facet of that training. Currently most training programs consist of an apprentice-type system in which the novice first learns to master the complex three-dimensional anatomy through reading, study of diagrams, CD-ROM material, lecture, and interactive three-dimensional models in some programs. Second, the trainee is exposed to clinical cases initially through observation in the operating theater, moving to assisting in procedures and finally to becoming the operator under the supervision of an attending physician. In the course of this exposure, skills training proceeds in the cadaveric lab, where actual human tissue is used for practice. With respect to ear surgery, this encompasses use of cadaveric temporal bones. As one might imagine, this type of cadaveric material is often difficult to obtain and carries the inherent risk of disease transmission. Each bone can only be used once. Reproducibility of resident performance on cadaveric bones is not high and makes generalizability of performance from one bone doubtful. Elstein, Shulman, and Sprafka (1978, chapter 4) have discussed possible reasons for this lack of reproducibility observed also in the medical problem solving.

Performance ratings or measurement in surgical skill has traditionally encompassed the subjective evaluation of an attending expert physician. The objective analysis of surgical skill is still somewhat in its infancy. As the need to reduce medical and surgical errors has become evident, the demand for more objective measures of skills performance has become paramount. The most widely known and accepted methodology employed is the objective structured assessment of technical skills (OSATS; Martin et. al., 1997; also see Faulkner, Regehr, Martin, & Reznick, 1996; Shah & Darzi, 2001). This methodology includes a section on step-by-step assessment of a surgical task followed by a section on global rating of performance. Each section may contain many individual elements, each requiring a score. The entire process is implemented with multiple observers to eliminate potential bias with rater/subject interaction. This is rarely implemented in institutions with any regularity.

We have recently attempted to develop and refine a type of final product analysis (FPA) tool for temporal bone dissection using a 35-item scoring instrument, the Welling Scale, Version 1 (WS1), which is used by expert raters to score end-product temporal cadaveric bone dissection (complete mastoidectomy with facial recess approach). Inter-and intrarater reliability of the WS1 has recently been demonstrated (Butler & Wiet, 2007). This study goes further in validating the use of the WS1 scores by analysis of scores that provide estimates of the relative contributions of sources of measurement error to the scores. We used the generalizability theory (GT) model of measurement to assess the characteristics of this measurement process and to determine the combinations of raters and temporal bones (scarce resources) that produce sufficiently small measurement error and adequate reliability (Brennan, 2000a, 2000b). The GT measurement model estimates the contribution of all sources of measurement error. This allows fine-tuning of the measurement protocol, so that measurement error contributions from the largest sources of error can be reduced sufficiently. In our study, one important source of measurement error comes from lack of agreement among raters. However, we show that this is a minor source of error as compared to inconsistent performance of residents across different temporal bones. Inconsistent performance across similar tasks has been reported in the measurement literature (Gross, 1994; Jarjoura, Early, & Androulakakis, 2004; Linn & Burton, 1994). We show how performance inconsistency can be reduced by appropriate choice of repeat performance and how total measurement can be reduced to a sufficiently small value by choice of rater and performance numbers.

Method

Institutional Review Board Approval

This project was performed with approval from the Institutional Review Board of the Ohio State University College of Medicine. Informed consent was obtained from all study subjects under the guidelines of the approved protocol.

The WS1

The 35 items scored in the WS1 (Butler & Wiet, 2007) are designed to measure various aspects of a quality end-product temporal bone dissection (a complete mastoidectomy with facial recess approach). The content validity of the 35 items is based on generally agreed-on features of a well-dissected temporal bone. There are nine subheadings that correlate to various key aspects of exposure and skeletonization within anatomical units of the dissected mastoid cavity. Within each subheading are items subjectively analyzed and scored in binary fashion as either adequately performed or not performed. Each of the 35 items covers an aspect of the dissection thought to be a key element in a quality dissection of that particular anatomical subunit. In the past, it was used as a grading instrument for end-product temporal bone dissection to score performance of postgraduate year (PGY) 3 residents in a temporal bone dissection course in the Department of Otolaryngology, College of Medicine, Ohio State University.

Procurement and randomization of temporal bones—Nineteen cadaveric bones were randomly selected from a collection of 30 temporal bones harvested from Ohio State University's Anatomical Donor Program after cadavers had been used for anatomy classes in the medical school. In addition to the human cadaveric bones, five Pettigrew plastic bones (see http://www.temporal-bone.com/) and one additional, previously drilled bone were added to the set, making a total of 25 bones rated. Twelve otolaryngology residents at Ohio State University, 3 in each of PGY 2, 3, 4, and 5, participated in this study. The PGY 2's had no prior temporal-bone drilling experience, the PGY 3's recently completed the annual temporal-bone dissection course, and PGY 4's and 5's completed the same course during their third year of training. Each resident was required to dissect two randomly assigned bones. Raters consisted of 4 neurotologists and 1 pediatric and 1 general otolaryngologist, who were blinded to PGY status of the dissector for each bone. A total of two rating sessions (Sessions 1 and 2) were performed for each bone 4 to 6 weeks apart. Five of the raters rated all students on both sessions, and one rater (Dr. Welling) rated all the students for one session only. The lapse between the two rating sessions was about 4 to 6 weeks. The bones were randomly re-ordered for the second rating session. Randomization procedures for bone assignment to residents of PGY 2, 3, 4, and 5, dissection procedures, and specifics of the rating methodology have been outlined in another publication (Butler & Wiet, 2007).

Statistical Analysis

All analyses were conducted using SAS Version 9.1 software. Data were summarized by either means for continuous variables or frequencies and percentages for categorical variables. Mixed models were used to estimate components of variance. The sources of measurement error (random effects) in the models are rater, bone, session, and their interactions. Repeat rating by the same rater allowed estimation of “session” variance components. Variance component estimates and corresponding 95% confidence intervals are reported. The method proposed in Brennan (2000b) was used to calculate the confidence intervals. In addition, generalizability coefficients were calculated for different number of raters and bones.

Results

Distributions by rater specialty, PGY year, bone type by PGY year, and bone difficulty by session type are described in Table 1. In Sessions 1 and 2, six and five experts rated the 25 bones, respectively. Each rater assessed bone difficulty (based on the inherent anatomy of an individual bone) on a 5-level, Likert-type scale. The reason for including this Likert-type scale for bone difficulty was to determine the impact of difficulty on measurement error. However, the bone difficulty was almost always judged as *average* or *moderate* (Difficulty Level 3): 92 out of 147 ratings and 91 out of 125 ratings in Sessions 1 and 2, respectively (see Table 1). In particular, although Rater 2 judged every bone as *average*, Rater 1 judged two bones as *very easy*, and Raters 3, 4, and 6 judged only a few bones as *difficult* (Session 1). Bones were randomly assigned to residents. No PGY 2's were assigned plastic bones and PGY 5's each received one plastic bone and one cadaveric bone (see Table 1).

Variance Components Models

Linear mixed models were used to identify different sources of variability across performance ratings. The first set of models was fitted using all bones (cadaveric and Pettigrew plastic bones). Initially, only random effects were included in the model: resident, rater, bone within resident (bone:resident), the interaction of Resident × Rater, and session variances (Model 1a). Resident represents the object of measurement, so this source is not considered part of the measurement error variance (MEV). The resident variance estimate is 107.37. All of the other components, excluding resident, add up to 147.84 (see Table 2, Model 1a, final row). Note that Model 1a included the main session effect and all the random effects associated with session (interactions with resident, rater, bone:resident, and Resident × Rater), whereas other models do not. The session effects tell us how reproducible ratings are from one rating session to the next, within rater (intrarater reliability). See Butler and Wiet (2007), which reports the kappa statistic analysis for inter- and intrarater reliability. Each variance component provides a different source of error or inconsistency of the same performance across sessions. From results of Model 1a (see Table 2, column 1), we found that session random effects (intrarater error) only contribute 5.7% of the total MEV. This means that raters are highly consistent from one session to the next (Butler & Wiet, 2007). Because session represents such a small component of MEV, we decided to focus on the larger and more critical components of measurement error.

For all the other models (1b, 1c, 1d), we used data from just the first session. In Model 1b, the resident variance estimate is 231.37. The addition of all the sources of measurement error (rater, bone:resident, Resident × Rater, and Rater × Bone:resident) is 137.05 (see Table 2, Model 1b, final row). The most important result from this model is that bone:resident represents a very large portion of the total MEV components, about 61% (84.08 out of 137.05).The bone:resident component reflects inconsistent performance by residents across the two bones. The interaction Resident × Rater (rater inconsistency in scoring each resident) explained only a very small part of MEV (about 10% = 13.81 out of 137.05; see Table 2).

When resident year (PGY) was included as a fixed effect in the model (see Table 2, Model 1c), the resident variance estimate was reduced from 231.37 to 91.34 (Models 1b and 1c).

But the estimates of the variance components of MEV remained unchanged. The bone:resident component was still the largest, about 61% of MEV (84.07 out of 137.04); whereas Resident × Rater explained the smallest fraction (about 10% = 13.81 out of 137.04; see Table 2, Model 1c). Thus, the partition of the MEV component is very similar in the two models (1b and 1c), but the variability because of resident was significantly reduced when PGY was taken into account. This is expected, as residents had different levels of training and varying levels of motivation to perform the required dissections. When, in addition to resident year (PGY), the inherent bone difficulty ratings were included in the model as fixed effects (see Table 2, Model 1d), the bone:resident component still represented about 62% of the MEV (81.25 out of 130.32). This suggests that bone difficulty scores failed to provide a valuable method for correcting for inconsistent performances across bones.

Analyses Using Only the Cadaveric Bones

We fitted a second set of models using only the cadaveric bones and data from the first session (see Table 3) because of complaints from the residents about the Pettigrew plastic bones being so different from normal cadaveric bone. Also, PGY 5 data were deleted from these analyses because PGY 5 residents drilled almost all plastic bones. Among the three PGY 5 residents, one drilled two plastic bones and the other two drilled one plastic bone each. Four models were fitted: random effects only (Model 2a); random effects and sequence (Model 2b); random effects and PGY (Model 2c); and random effects, PGY, and bone difficulty (Model 2d).

With removal of the plastic bones (and PGY 5's), we were able to estimate a sequence effect in Model 2b, to determine whether residents consistently performed better or worse on the second cadaveric bone that they were randomly assigned. In Model 2b, we found that sequence effect was not significant ($p = .09$), and the bone:resident variance could not be explained by a sequence effect. The bone:resident variance with sequence in this model (Model 2b) is 52.71, and the bone:resident variance is 58.31 in Model 2a. The least squares mean percentage scores were larger for bones drilled in first place (43.81 for the first bone and 36.22 for the second).

The resident variance estimate was 309.83 (see Table 3, Model 2a). When sequence was included in the model, the resident component was reduced to 201.31 (Model 2b); and when PGY was included in the model, the resident component was reduced to 89.98 (Model 2c). This difference in PGY performance is clearly seen by a comparison of pre- to posttraining residents (PGY 2 with 3 residents and 6 bones versus PGY 3–4 with 6 residents and 11 bones; $p = .0014$), for which the difference was about 27 percentage points (26.90% correct in pretraining and 54.11% in posttraining). Only minor changes in the distribution of the variance components were observed when bone difficulty was included in this model (see Table 3, Model 2d). Similarly to Models 1a through 1d, the bone:resident component always represented the largest portion of MEV (between 60% to 62% in Models 1a through 1d and between 44% to 51% in models 2a through 2d). When bone difficulty was included in Models 1d and 2d, this variance component was changed by only 3% or 2%, respectively (from 84.07 to 81.26 in Table 2 and from 58.91 to 60.35 in Table 3).

Generalizability Coefficient (GC)

GT stresses the importance of multiple sources of measurement error and provides methods for configuring a measurement process that produces reliable scores. Generalizability coefficients can be used to determine the ideal number of raters and bones—in this case, to produce reliable scores (Brennan, 2000a, 2000b).

The GC is the percentage of variance in resident scores produced by a chosen measurement procedure that represents the true difference among residents' performance. This is calculated as follows:

$$\text{GC}=\frac{\text{Resident Variance Component}}{\text{Resident Variance Componetnt}+\text{MEV}}$$

Values greater than 0.80 are considered adequate for most standardized measurement instruments, but lower values (about 0.50) are often found in performance instruments when subjects are homogenously competent performers—in this case, it means that resident variance is low.

GC values were calculated for scores produced by averaging across different numbers of raters and bones (see Table 4), using results from Model 2c. The first row in Table 4 represents the study conditions (six raters and two bones per resident). The GC's were larger when only random effects were included in the models compared to when the fixed effect “PGY level” was included in the models (results not shown). This is because of the reduction in true variance across residents when this effect is taken into account. The GC's ranged from 0.56 to 0.71. The standard error (*SE*) ranged from 6.05 to 8.34. GC's increase when the number of bones and/or number of raters increases. When six raters and two bones (study scenario) were used, the GC was 0.71. When six raters and one bone were used, the GC was reduced to 0.56. The same GC as the one in the study scenario was obtained when only two raters and three bones were used (first and final rows of Table 4).

When faced with highly homogeneous performance across subjects, the *SE* of measurement is often the focus rather than GC (Jarjoura et al., 2004; Linn & Burton, 1994). GC depends on subject variance, and high values imply good discrimination among them. In contrast, identification of incompetent performance, relative to a standard, mainly requires consideration of the size of the SEM. In this study, two raters and three bones produces an SEM of 6.14, which is 14% of the mean performance (i.e., CV is 0.14; see Table 4). Whether a CV of 0.14 is small enough to identify incompetence depends on knowledge of the gap between incompetent and competent performers. Recall that the difference between PGY2 residents who had no training in these surgeries and the others was 27 points, which represents 4.4 SEM's. This difference between pretraining (PGY 2) versus posttraining (PGY 3 and 4) was significant (p = .0014). In other words, the separation between the means is large relative to measurement error, indicating that the instrument is sensitive enough to easily differentiate these two levels of experience.

Discussion

Use of GT has allowed us to analyze various components of measurement error to better understand the properties and results of the WS1. In addition, it has allowed us to develop a protocol for administration of the WS1 that provides for the maximum use of limited resources (raters and temporal bones) that provide sufficiently small measurement error. This methodology is important when one considers that development of additional OSATS testing will continue to be needed as performance measurements become more important in credentialing. Objective performance measurement is quickly becoming an important component of surgical education and maintenance of certification as public demands mount for reduction in error and demonstration of improved outcomes. “Pay for performance” concepts are being studied and beginning to be implemented by the insurance industry and government regulating bodies. Because surgical performance has been difficult to measure objectively in the past, development of measurement tools that are valid and applied in a thoughtful manner is paramount to providing accurate assessment (Zirkle, Taplin, Anthony, & Dubrowski, 2007). In Zirkle et al. (2007), three different assessment measures of performance in the temporal bone laboratory were studied and compared between two groups of residents' expertise (novice and experienced) evaluated by only two raters. These three measures were the Global Rating Scale, Task-Based Checklist, and a final product-analysis measure. The first two are used by the OSATS. They looked at interobserver correlations and concluded that the measures were reliable because correlations were between .73 and .81. They then compared the residents' totals scores (average across raters) from the three measures between the two groups (novice or experienced) by the Mann–Whitney U test and concluded that the two OSATS instruments were “objective measures” because the two groups were significantly different in terms of total scores. They then fitted logistic regression models to determine whether any of the three scales was a predictor of expert opinion (EO). They found that one of the OSATS measures was a good predictor of EO, based on high pseudo-R^2 values. This analysis partitioned the problem into three questions. We believe that these results may not properly address the measurement properties of these instruments. Our concerns are the following:

1. Correlation between two raters. The correlation between the two raters could be high, but the total scores could be very different (one rater could have consistently assigned lower scores than the other rater). So the correlation between the two raters is not an indicator of accuracy.

2. Differences in total scores between the two groups (novice and experienced) do not indicate that the instrument provides sufficiently accurate discrimination among individuals' performances.

3. Associations between the three scales and expert opinion showed no significant association for these measures. Finally, logistic regression pseudo-R^2 values are not indicators of goodness of fit; thus, the use of pseudo-R^2 is not recommended to validate models (Hosmer & Lemeshow, 2000, p. 164, section 5.2.2).

There are several sources of measurement error involved in the evaluation process that were completely ignored by the way that these authors performed the analysis.

> Use of GT is one way in which we can validate our performance metrics as we move forward in developing assessment tools. This study demonstrates the utility of such analysis in general and, specifically, with respect to use of the WS1 for assessment of temporal-bone dissection performance.

In our study, bone:resident was always the largest component of MEV. This means that residents performed inconsistently across the two bones that they were assigned, and this inconsistency was larger than other sources of measurement error. This is a common observation in different performance testing environments (Elstein et al., 1978; Gross, 1994; Jarjoura et al., 2004; Linn & Burton, 1994). Why this occurs is a matter of speculation. In our study, residents may have shown performance variability depending on their state of mind, unique features of each of the temporal bones assigned (beyond the difficulty rating), previous sleep, and problematic daily variations. These types of interactions require further study as one seeks to develop accurate performance measures and must be taken into account as one develops testing protocols.

In the analysis of error because of rating sessions (Rating Sessions 1 and 2), we found that repeated ratings of the bones did not show inconsistencies nearly as large as other sources of measurement error. The rating session variance component represented a small portion of MEV, suggesting that only one rating session per rater is needed to obtain sufficiently accurate measurements (i.e., raters do not need to reevaluate the bones). This is consistent with the high intrarater kappa obtained when measuring rater agreement for the WS1 (Butler

& Wiet, 2007). PGY level was highly significant and explains two-thirds of the true variance in resident performance. Including PGY level in the model did not change the partition of the error variance components. The difference in PGY performance between pre- and posttraining was large (4.4 SEM's) and significant. This provides evidence of validity of the instrument (i.e., the large gap in performance indicates that the WS1 is sensitive to training. In particular, the bone:resident component did not change and remained as the largest component of MEV. Thus, we conclude that the most efficient way to reduce MEV and thereby better discriminate among performances is by increasing the number of bones (i.e., performances) that each resident is rated on. In fact, because bone:resident variance is so large, it is not possible to sufficiently reduce measurement error without using multiple bones. Furthermore, when only two raters and three bones per resident were used, the same GC's values as those with six raters and two bones were obtained. Thus, it is more important to increase the number of bones than the number of raters to evaluate residents' performances using the WS1. Each resident was asked to perform the same surgical procedure, so the variation observed in the way the bones were drilled is mostly caused by variation in bone structures. Thus, this would not be considered as a “case-specificity” problem. We obtained consistently high GC's under numerous testing scenarios and identified that two raters grading three bones, when PGY level was taken into account, yielded a GC of 0.70 and CV = 0.14; the same values were obtained when using six raters grading two bones. This testing structure, however, may not be feasible for otololaryngology residency programs, which may have limited access to human cadaveric temporal bones or limited raters. If two raters and two bones are used (a more practical testing condition), the GC reduces to 0.64 and CV increases to 0.16.

Because Pettigrew plastic bones are inherently different than cadaveric bones, we further assessed the cadaveric bones without including the plastic bones in the model (Model 2a through 2d). Although the plastic bones were not rated as more difficult, we found that the bone:resident component dropped from 84.08 to 58.31 (from about 60% to 48% of the MEV) in the random models (Model 1b through Model 2a). A similar trend was observed in the models with PGY level controlled. This suggests that the Pettigrew plastic bones create substantial error in performances, such that the inconsistency between performances on plastic and cadaveric bones is very large. (There was a total of five plastic bones, and these were randomly assigned as follows: one to a PGY 3; one to a PGY 4, and three to each of the PGY 5's; see Table 1). We concluded that plastic bones are not as useful as cadaveric bones for rating performance using the WS1. This was the first time that residents dissected Pettigrew plastic bones, and we did not assess whether surgical practice with plastic bones would have increased either cadaveric bone dissection scores or plastic-only bone scores. Although all otolaryngologists are trained in temporal bone surgery, neurotologists are considered the more expert based on their additional training and practice limited to the temporal bone and related nervous system structures.

In a sensitivity analysis (not reported in detail), we looked at the variability of percentage scores because of rater specialty by comparing the four neurotologists with the other raters (one pediatric and one general otolaryngologist). The neurotologists discriminated among the residents much better than all other raters. The resident variance component was 320 for neurotologists, whereas for the others, it was only 27. In addition, for neurotologists, the Resident × Rater variance component was very small (0.24), whereas the same component was 30 for other raters. This suggests that neurotologists were very consistent in their ratings of each of the residents.

A major limitation of this study is the small number of residents for estimating the resident variance component. This is reflected in the wide confidence intervals in Tables 2 and 3. The number of cadaveric bones per resident was usually two, and these were used for estimating the large MEV component (bone:resident). The confidence interval for this component was also large. However, our conclusion that represents a major component of inconsistency in performance is well supported by the lower limit of the confidence interval (note that the lower limit is still a large component of the total MEV: between 27% and 31% in Table 2 and between 20% and 23% in Table 3). But this limitation is inherited in any statistical modeling of data from one residency program for one measurement procedure. We plan to obtain similar data from multiple sites to more accurately estimate these components. In addition, our estimates of the standard error of measurement for different combinations of numbers of raters and bones (see Table 4) are fairly accurate. For example, for two bones and two raters, the *SE* of the MEV was 17.43 (using the variance–covariance matrix of the variance components for Model 2c).

One of the limitations of grading cadaveric bones is that each bone and its inherent difficulty are unique. We expected that we could correct for varying levels of difficulty of the bones by asking raters to provide difficulty assessments. It turned out that bone difficulty had almost no impact on reducing the main error variance components for bones (bone:resident). When bone difficulty was included in the models, this variance component was reduced or increased by approximately 3%.

In our particular case, using GT to measure components of error in our testing procedure has helped us to determine the best allocation of scarce resources (cadaveric material and expert raters). Other areas or surgical performance testing that require scarce resources to implement testing protocols should consider GT to help address the optimal allocation of those resources to provide the best measurement protocol. As an example, knowing the least number of bones needed to accurately assess performance is useful information for planning additional studies that require a measurement of temporal bone dissection performance. Also, knowing the minimal number of expert raters needed to accurately assess performance using the WS1 is also necessary to make the process feasible. Our ultimate goal is to develop an accurate tool for identifying incompetent performances as well as the methodology to implement that tool in a reasonable fashion in different training scenarios. In other words, use of the WS1 in the format outlined above could be implemented as a reliable measurement of performance of technical skills in assessing a resident's readiness to perform surgery on patients. In addition, use of the WS1 can also valuable in comparing the efficacy of different training methodologies such as computer-based simulation environments versus traditional training (Wiet et al., 2002).

In summary, when using GT, we determined that the largest source of measurement error occurred because of residents' inconsistent performance across bones. Raters' disagreement introduced only small error into scores. The WS1 with two raters and two bones for each study subject provides adequately small measurement error.

Acknowledgments

This work was supported by the National Institutes of Health (NIDCD RO1 DC006458-01).

Appendix

Welling Scale for Temporal Bone Dissection (WS1)

Please grade each item. 0 = *incomplete, inadequate dissection*, 1 = *complete, adequate*
Cortex

1.	Cortex rounded at linea temporalis	0	1
2.	Cortex rounded from linea temporalis to middle cranial fossa	0	1
3.	Thinning of posterior canal wall	0	1
4.	Complete saucerization Tegmen mastoideum	0	1
5.	Dissection parallels curve of the dura	0	1
6.	Completely exposed	0	1
7.	No holes	0	1
8.	No cells remain Sigmoid Sinus	0	1
9.	No holes	0	1
10.	No cells	0	1
11.	No overhang	0	1
Sinodural Angle			
12.	Sharp	0	1
13.	No cells remaining Digastric Ridge	0	1
14.	Identified	0	1
15.	Digastric tendon followed to stylomastoid foramen External Auditory Canal	0	1
16.	Canal Wall Up	0	1
17.	Without holes	0	1
18.	Without cells Semicircular Canals Skeletonized	0	1
19.	Horizontal	0	1
20.	Superior	0	1
21.	Posterior	0	1
22.	Blue lined without fenestra Facial Nerve	0	1
23.	Identification of nerve at the stylomastoid foramen	0	1
24.	Identification of nerve at the external genu	0	1
25.	Identification tympanic segment	0	1
26.	Identification of nerve at cochleariform process	0	1
27.	No exposed nerve sheath	0	1
28.	Identification of chorda tympani or stump	0	1
29.	Facial recess completely exposed Additional Anatomical Structures	0	1
30.	Stapedial muscle dissected	0	1
31.	ELS transition to duct	0	1
32.	Blue line of basal turn of the cochlea	0	1
33.	Identification of carotid artery in middle ear	0	1
34.	Identification of jugular bulb	0	1
35.	Skeletonization of jugular bulb	0	1

Difficulty of temporal bone					
	1	2	3	4	5
	Very Easy	Easy	Average	Difficult	Very Difficult

References

Brennan RL. Performance assessments from the perspective of generalizability theory. Applied Psychological Measurement. 2000a; 24:339–353.

Brennan, RL. Generalizability theory. Springer; New York: 2000b.

Butler NN, Wiet GJ. Reliability of the Welling Scale (WS1) for rating temporal bone dissection performance. Laryngoscope. Oct; 2007 117(10):1803–1808. [PubMed: 17721407]

Elstein, AS.; Shulman, LS.; Sprafka, SA. Medical problem solving: An analysis of clinical reasoning. Harvard University Press; Cambridge, MA: 1978.

Faulkner H, Regehr G, Martin J, Reznick R. Validation of an objective structured assessment of technical skill for surgical residents. Academic Medicine. 1996; 71:1363–1365. [PubMed: 9114900]

Gross LJ. Inter-rater reliability reconsidered: Performance assessment using one examiner per candidate. Evaluation & the Health Professions. 1994; 17:465–484.

Hosmer, DW.; Lemeshow, S. Applied logistic regression. 2nd ed.. John Wiley; New York: 2000.

Jarjoura D, Early L, Androulakakis V. A multivariate generalizability model for clinical skills assessments. Educational and Psychological Measurement. 2004; 64(1):22–39.

Linn R, Burton E. Performance-based assessments: Implications of task specificity. Educational Measurement: Issues and Practice. 1994; 13:5–8.

Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. British Journal of Surgery. 1997; 84(2):273–278. [PubMed: 9052454]

SAS Institute Inc.. SAS. version 9.1. Author; Cary, NC: 2008.

Shah J, Darzi A. Surgical skills assessment: An ongoing debate. BJU International. 2001; 88:655–660. [PubMed: 11890231]

Wiet GJ, Stredney D, Sessana D, Bryan J, Welling DB, Schmalbrock P. Virtual temporal bone dissection: An interactive surgical simulator. Otolaryngology–Head and Neck Surgery. 2002; 127:79–83. [PubMed: 12161735]

Zirkle M, Taplin MA, Anthony R, Dubrowski A. Objective assessment of temporal bone drilling skills. Annals of Otology, Rhinology and Laryngology. 2007; 116(11):793–798.

Table 1

Distribution by Rater Specialty, Bone Type by PGY Year, and PGY Year by Bone Difficulty for Each of the Sessions

	N
Rater specialty	
Neurotologists (raters 3,5,6)	2
Neurotology fellow (rater 2)	1
Pediatric otolaryngologist (rater 1)	1
General otolaryngologist (rater 4)	1

Bone type distribution by PGY year (*N* = 25)	**Cadaveric (*N* = 20)**	**Plastic (*N* = 5)**
PGY 2	6	0
PGY 3	6	1
PGY 4	5	1
PGY 5	3	3

PGY by bone difficulty (Session 1; *N* = 150)[a]	**Difficulty Level**		
	1,2,3,4,5		
PGY2	8	25	3
PGY3	9	27	6
PGY4	9	19	6
PGY5	4	21	10
Total	30	92	25

PGY by bone difficulty (Session 2; *N* = 125)	**Difficulty Level**		
	1,2,3,4,5		
PGY2	2	26	2
PGY3	3	27	5
PGY4	7	19	4
PGY5	2	19	9
Total	14	91	20

Note: PGY = postgraduate year.

[a] Three values were missins.

Table 2

Variance Components Estimates (VCE), Percentages of Measurement Error Variability (MEV), and 95% Confidence Intervals (CI) Using Cadaveric and Plastic Temporal Bones

	Model 1a With Session		Model 1b Random		Model 1c With PGY Level		Model 1d With PGY Level and Bone Difficulty	
Random Effects	**VCE**	**%**	**VCE**	**%**	**VCE**	**%**	**VCE**	**%**
Resident	107.37		231.37		91.34		99.91	
CI	48.96, 394.22		105.50, 849.49		41.65, 335.38		45.56, 366.82	
Rater		0	5.87	4.28	5.87	4.28	3.79	2.91
CI		—	2.68, 21.54		2.67, 21.54		1.73, 13.92	
Bone:resident	88.79	60.06	84.08	61.35	84.07	61.35	81.25	62.35
CI	40.49, 325.99		42.70, 235.48		42.70, 235.48		40.76, 234.31	
Resident × Rater	10.01	6.78	13.81	10.07	13.81	10.07	15.09	11.58
CI	5.12, 38.44		6.63, 44.19		6.64, 44.19		7.57, 43.54	
Rater × Bone:resident	40.55	27.43	33.29	24.29	33.29	24.29	30.19	23.17
CI	32.94, 51.14		24.27, 48.53		24.27, 48.53		21.73, 44.88	
Session components		5.74						
Session × Resident	3.09							
CI	1.41, 11.36							
Session × Rater	4.68							
CI	2.14, 17.20							
Session × Bone:resident	0.72							
CI	0.33, 2.63							
Session × Resident × Rater	0.00							
CI	—							
Total of MEV components	147.84		137.05		137.04		130.02	

Note: PGY = postgraduate year. The factors constitute the measurement error components; the last row is the addition of all the estimates except the “resident” component.

Table 3

Variance Components Estimates (VCE), 95% Confidence Intervals (CI), and Percentages of Measurement Error Variance (MEV) Using Only Cadaveric Bones and PGY Levels 2, 3, and 4.

	Model 2a Random		Model 2b With Sequence		Model 2c With PGY Level		Model 2d With PGY Level and Bone Difficulty	
Random Effects	**VCE**	**%**	**VCE**	**%**	**VCE**	**%**	**VCE**	**%**
Resident	309.83		201.31		89.98		96.67	
CI	141.28, 1137.57		91.80, 739.13		41.03, 330.39		44.08, 354.95	
Rater	10.50	8.58	10.67	8.8	10.50	8.50	5.29	4.50
CI	4.79, 38.54		4.86, 39.16		4.79, 38.54		2.41, 19.42	
Bone:resident	58.31	47.69	52.71	43.75	58.91	47.95	61.14	43.30
CI	26.59, 214.10		24.04, 193.54		26.86, 216.30		27.88, 224.48	
Resident × Rater	11.75	9.61	18.41	15.28	11.75	9.56	17.16	14.50
CI	5.36, 43.12		8.40, 67.60		5.36, 43.12		7.83, 63.02	
Rater × Bone:resident	41.70	34.11	38.68	32.11	41.70	33.94	34.92	29.46
CI	28.17, 68.07		33.17, 45.70		28.17, 68.07		23.16, 58.62	
Total of MEV components		126.26		120.47		122.86		118.51

Table 4

Generalizability Coefficients (GC) for Different Number of Raters and Bones; and Standard Error (*SE*) of Measurement for Model 2c

No. of Raters	No. of Bones	*SE* of Measurement	CV	GC
6	2	6.05	0.13	0.71
6	1	8.34	0.18	0.56
2	2	7.14	0.16	0.64
1	2	7.84	0.17	0.59
2	3	6.14	0.14	0.70