Comparison of Summative Temporal Bone Dissection Scales Demonstrate Equivalence

Jordan B. Hochman[1] Justyn Pisa[2] Shubhi Singh[2] Michael Gousseau[3] Bert Unger[1,4]

[1] Division of Neurotologic Surgery, Department of Otolaryngology Head and Neck Surgery, Faculty of Health Sciences, University of Manitoba, Manitoba, Canada
[2] Department of Otolaryngology Head and Neck Surgery, Health Sciences Centre, Winnipeg, Manitoba, Canada
[3] Department of General Otolaryngology, Dr. Michael Gousseau Medical Corporation, Portage La Prairie, Manitoba, Canada
[4] Laboratory for Surgical Modeling, Simulation and Robotics, University of Manitoba, Manitoba, Canada

Address for correspondence Justyn Pisa, AuD, Department of Otolaryngology - Head and Neck Surgery, Health Sciences Centre, GB421, 820 Sherbrook Street, Winnipeg , Manitoba R3A 1R9, Canada (e-mail: jpisa@hsc.mb.ca).

Int Arch Otorhinolaryngol 2022;26(4):e556–e560.

Abstract

Introduction Temporal bone surgery is a unique and complicated surgical skill that requires extensive training. There is an educational requirement to maximize trainee experience and provide effective feedback.

Objective We evaluate three temporal bone dissection scales for efficacy, reliability, and accuracy in identifying resident skill during temporal bone surgery.

Methods Residents of various skill levels performed a mastoidectomy with posterior tympanotomy on identic 3D-printed temporal bone models. Four blinded otologic surgeons evaluated each specimen at two separate intervals using three separate dissection scales: the Welling Scale (WS), the Iowa Temporal Bone Assessment Tool (ITBAT), and the CanadaWest Scale (CWS). Scores from each scale were compared in their ability to accurately separate residents by skill level, inter- and intrarater reliability, and efficiency in application.

Results Nineteen residents from 9 postgraduate programs participated. Assessment was clustered into junior (postgraduate year or PGY 1, 2), intermediate (PGY 3) and senior resident (PGY 4, 5) cohorts. Analysis of variance (ANOVA) found significant differences between cohort performance ($p < 0.05$) for all 3 scales considering the PGY level and the subjective account of temporal bone surgical experience. The inter-rater reliability was consistent across each scale. The intrarater reliability was comparable between the CWS (0.711) and the WS (0.713), but not the ITBAT (0.289). Time (in seconds) to complete scoring for each scale was also comparable between the CWS (42.7 ± 16.8), the WS (76.6 ± 14.5), and the ITBAT (105.6 ± 38.9).

Conclusion All three scales demonstrated construct validity and consistency in performance, and consideration should be given to judicious use in training.

Keywords
- temporal
- bone
- training
- dissection
- scales

received
February 22, 2021
accepted after revision
September 11, 2021
published online
January 28, 2022

DOI https://doi.org/10.1055/s-0041-1740162.
ISSN 1809-9777.

Thieme Revinter Publicações Ltda., Rua do Matoso 170, Rio de Janeiro, RJ, CEP 20270-135, Brazil

Introduction

Temporal bone surgery is complex. Surgical trainees must have a strong understanding of anatomical relationships and proficiency with equipment to be safe. Historically, emphasis was placed on structured training with cadaveric specimen. Excitingly, there is now an expansive number of educational media that can be employed, including virtual reality (VR), 3D-printed, and augmented reality.[1–9]

The benefits of resident education using 3D-printed models include their physical similarities to cadaveric bone, relatively low-cost, ease of acquisition, and availability for presurgical rehearsal.[10] While a different modality, VR temporal bone models offer residents another low-cost and interactive opportunity for surgical training without risk to patients.[11] However, while a considerable effort has been put toward the development of these educational resources, there has been less attention on resident skill assessment.[12,13]

These tools are needed, as the educational paradigm no longer permits deliberate practice, owing to resident and facility availability.[13] In the current context, trainee skill development is challenged, juxtaposed with increasing scrutiny on outcomes. Furthermore, there is a trend toward competence by design (CBD), heightening the need for objective skill assessment for advancement, including accreditation.

Several formative assessment tools have been developed.[14,15] These tools are not being reviewed, as they represent a different process in which the evaluator needs to be present for the dissection. While there is great value in formative assessment, it is a more time intensive process. Rather, the present paper will focus on validated summative temporal bone dissection scales that evaluate a final dissection product.

With competency-based medical education, accurate assessment of technical competence assumes greater importance. An effective summative assessment tool needs to be easy in application, accentuate success, and adequately differentiate between safe and possibly injurious and absolute deleterious activities.[13]

There are several validated temporal bone grading tools.[16,17] The schema with the longest clinical use is the Welling Scale from Ohio State University (WS-1).[16] We will further consider the more recently developed CanadaWest Scale (CWS) and the Iowa Temporal Bone Assessment Tool (ITBAT).[12,18,19]

Objective

The objective of the present study is to compare the three tools by measuring trainee performance on a canal wall-up mastoidectomy with posterior tympanotomy, a minimum standard skill, performed on identic 3D printed models. The outcome of interest is the ability of the scales to differentiate trainee skill, speed in application of the tools, and to quantify intra- and inter-rater reliability.

Methods

Institutional Research Ethics Board approval was obtained prior to initiation of the study (HS18582). Residents attending a Canadian Otolaryngology-Head and Neck Surgery (CSOHNS) meeting were invited to participate. Participation was voluntary and resident surgeons registered for the study preceding the conference. Participant accrual was a sample of convenience. Each participant was provided a unique identifier and all data was assessed in a blinded fashion.

Participants were provided with an identical 3D-printed temporal bone and performed a canal wall up mastoidectomy with posterior tympanotomy with an otic drill (Stryker, Kalamazoo, Michigan, USA) and operating microscope (Zeiss, Jena, Germany) as per convention. Demographic data was obtained from the registration of the participants (►**Table 1**).

The process for generating the model has been previously published.[1–6,9] Volumetric computerized tomography (CT) images are segmented into anatomical regions of interest. These meshes are combined, voxelated and sliced into sections for printing, after which alignment fiducially are added. Individual slices are then combined to produce a final physical model (►**Fig. 1**).

The completed specimens were graded by four independent blinded otologists from a single institution, using the WS-1, the ITBAT and the CW scales. The specimens were graded at 0 and then at 4 weeks following the initial scoring. Assessment duration was determined for each scoring session.

The Welling Scale was modified (WS-1) to focus on canal wall up mastoidectomy with posterior tympanotomy with the scale reduced from 35 to 21 assessment measures.

Table 1 Demographics of the Participants

Demographics of the participants, $n = 19$			
Gender	Female $n = 8$ (42%)	Male $n = 11$ (58%)	
Postgraduate year	Junior (PGY1,2) $n = 4$ (21%)	Intermediate (PGY3) $n = 9$ (47%)	Senior (PGY4,5) $n = 6$ (32%)
Perceived level of experience	None/Little $n = 7$ (37%)	Some $n = 9$ (47%)	Substantial $n = 3$ (16%)

Abbreviation: PGY, postgraduate year.

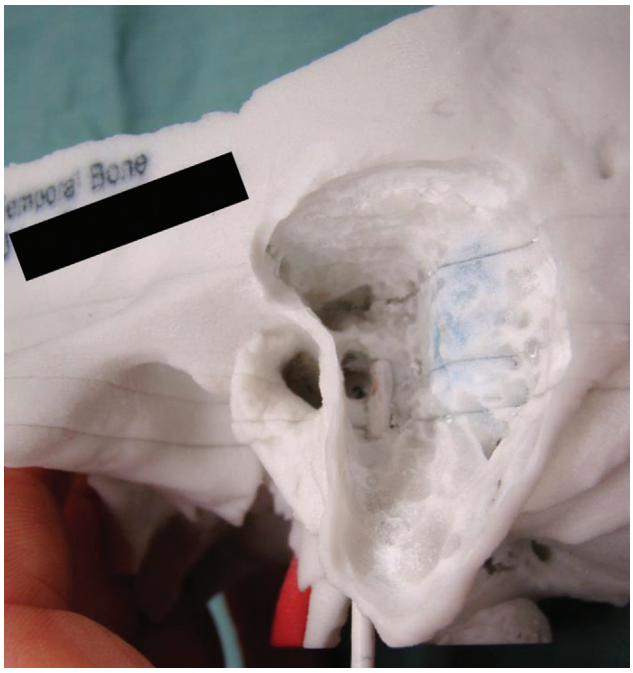

Fig. 1 Dissected temporal bone specimen. Depiction of a dissected temporal bone model. The carotid artery, sigmoid sinus, and facial nerve are evident.

Conversely, the ITBAT contained 23 assessment items, while the CW scale included 12 items.

For the purposes of analysis, participants were grouped by experience level, where PGY1–2 residents were included in the "junior" cohort, PGY3 residents were deemed to be of "intermediate" experience level, and PGY4–5 residents were combined into a "senior" group. A disparate analysis grouped participants by perceived experience with temporal bone surgery. Those who believed they had "little" experience were grouped separately from those who felt they had "some" experience with the procedure and from those who felt they had "substantial" experience.

Both postgraduate level and perceived level of experience with the canal wall up mastoidectomy with posterior tympanotomy were assessed.

An ANOVA analysis was used to compare the average scores between the scales. The Cohen kappa and intraclass correlation coefficient (ICC) were conducted to assess intra- and inter-rater reliability.

Results

Nineteen residents participated in the present study, with demographic information presented in ►**Table 1**. The average scores for each scale when considering the PGY level are displayed in ►**Fig. 2**. Each scale illustrated statistical significance between the junior and intermediate cohorts ($p < 0.001$) and the junior and senior cohorts ($p < 0.001$). However, none of the tools found a statistically significant result between the intermediate and senior cohorts.

The ability of each scale to differentiate by perceived experience level can be seen in ►**Fig. 3**. In each cohort, all 3 schemas showed a statistically significant result between the "little" and "some" experience categories ($p < 0.001$) and the "little" and "substantial" cohorts ($p < 0.001$). There was further significance found between the "some" and the "substantial" cohorts for both the ITBAT and CWS but not for the WS-1.

In assessing the reliability of scoring between the 4 raters using a 2-way random effects ICC, a high degree of reliability

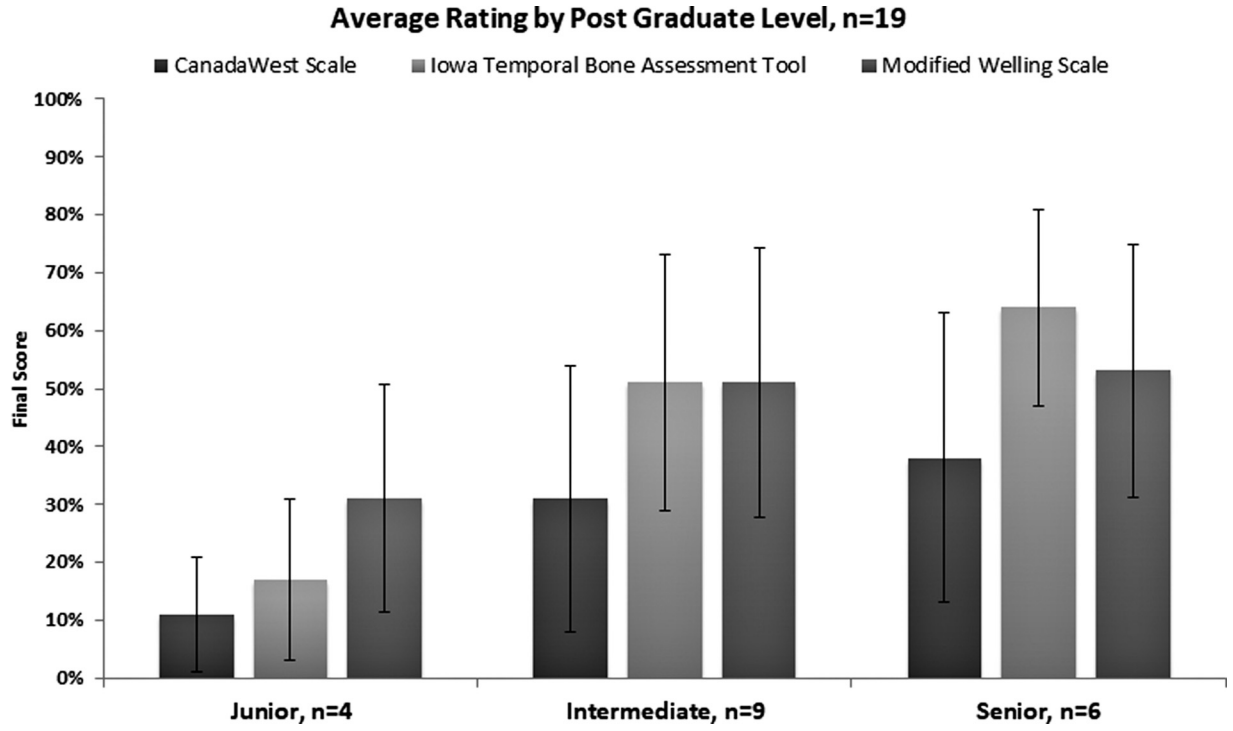

Fig. 2 Cohort performance by postgraduate level, converted to percentage score relative to each scale.

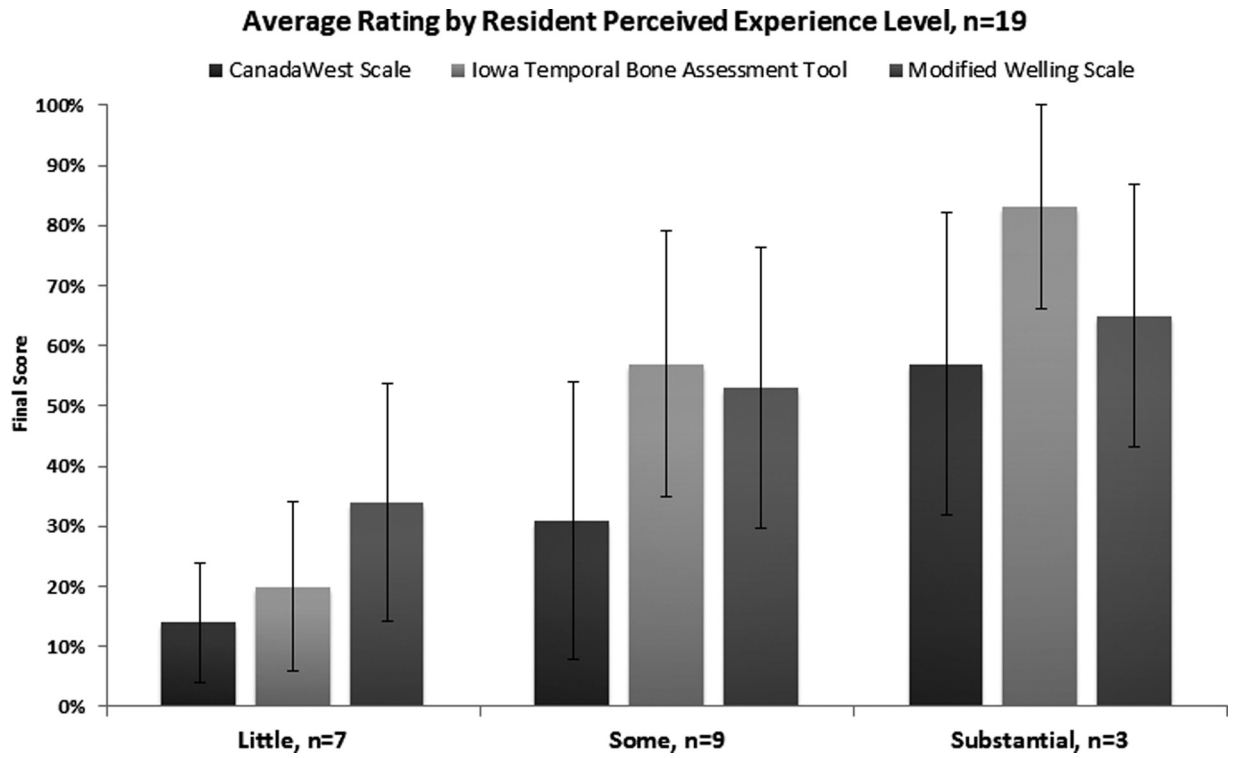

Fig. 3 Cohort performance by perceived experience level, converted to percentage score relative to each scale.

of 0.862 (95% confidence interval [CI]: $0.788 < ICC < 0.918$) was found for all 3 scales. The Fleiss Kappa inter-rater reliability (►**Table 2**) for the WS-1 was 0.917, 0.858 for the CW scale, and 0.790 for the ITBAT. A mean Cohen kappa for intrarater reliability for the modified WS-1 was 0.713, 0.711 for the CW scale, and 0.289 for the ITBAT.

The expert surgeons were timed during the scoring procedure. The average time needed for scoring the ITBAT was 105.6 (±38.9) seconds, 76.6 (±14.5) seconds for the WS-1, and 42.7 (±16.8) seconds for the CWS. In comparing the scales, the WS-1 is significantly shorter than the ITBAT ($p < 0.05$). The scoring duration of the CWS is significantly shorter than both that of the ITBAT and of the WS-1 ($p < 0.05$).

Discussion

Temporal bone dissection is unique in surgery. The structures of interest are encased in bone, amplifying complexity. With the recent evolution toward competency-based education, progress is required in both deliberate practice and evaluation of these opportunities. The summative measures reviewed in the present paper can delineate capacity and begin an objective conversation with a resident toward improving skills.

Table 2 Intra- and inter-rater reliability for the WS-1, the ITBAT and the CW scale

Reliability Measures, $n = 19$			
Intrarater reliability	ITBAT	CWS	WS-1
	0.289	0.711	0.713
Inter-rater reliability	ITBAT	CWS	WS-1
	0.790	0.858	0.917

Abbreviations: CWS, CanadaWest scale; ITBAT, Iowa Temporal Bone Assessment Tool; WS-1, Welling scale.

All three scales demonstrated value in improving performance across PGY and perceived experience. It should be highlighted that the scales were able to differentiate all three cohorts by perceived level of experience with the task. This may be a more apt consideration than PGY. The absence of significance between the intermediate and senior cohorts may be a function of the fidelity of each tool to discriminate performance, or it could also be the result of a strong intermediate cohort performance. Furthermore, it is always more complicated to distinguish near-expert performance due to a plateauing learning curve.[12,18]

It should also be noted that the assessment tools used in grading the final product all rely on Likert scales for grading and may have difficulty distinguishing fine gradations in technique. This ceiling effect may reduce the ability to distinguish between participants in the later stages of training.

Due to the relatively small sample size ($n = 19$), the limited statistical power of the present study (70.35%) may have confounded the ability to find statistically significant differences between participant cohorts. A post hoc analysis found that an n of $\sim$ 26 participants would be required to achieve statistical power at the 80% level.

All scales showed strong inter-rater reliability. However, intrarater scores were more modest. It is difficult to account for this outcome. A contributing feature is probably the absence of training for the expert evaluators. There is

bidirectional literature, both supporting the use of graders who are unfamiliar with a task to be evaluated as well support for training the experts.[3,19] In the present study neither condition was provided. Experts were permitted to make employ of the scales at their discretion. The significance of this approach is difficult to quantify.

A significant strength of the present study is the use of a printed temporal bone model, which eliminates the confounding cadaveric variability across participants.

Limitations include the small sample size and could be addressed in a subsequent study with more participants. Furthermore, as delineated, while the assessors are expert surgeons, they were not trained.

Conclusion

The WS-1, the ITBAT, and the CWS demonstrate strong performance equivalence and are easy to execute. These summative dissection tools may be employed as a component of global otologic education.

Declarations
Study supported by:
1. Stryker Canada
2. Dept. of Otolaryngology Head and Neck Surgery, University of Manitoba

Disclosure/Funding
Annual research funding from Advanced Bionics. No direct financial support was provided for this research.

Conflict of Interests
The authors have no conflict of interests to declare.

Acknowledgments
Darren Leitao, MD; Jodi Jones, MD; Brian Blakley, MD.

References

1 Unger BJ, Hochman JB, Kraut J. Method and system for rapid prototyping of complex structures. U.S. Patent Application 13/935,681, filed January 30, 2014
2 Hochman JB, Kraut J, Kazmerik K, Unger BJ. Generation of a 3D printed temporal bone model with internal fidelity and validation of the mechanical construct. Otolaryngol Head Neck Surg 2014; 150(03):448–454
3 Wong D, Unger B, Kraut J, Pisa J, Rhodes C, Hochman JB. Comparison of cadaveric and isomorphic virtual haptic simulation in temporal bone training. J Otolaryngol Head Neck Surg 2014;43 (01):31
4 Hochman JB, Rhodes C, Wong D, Kraut J, Pisa J, Unger B. Comparison of cadaveric and isomorphic three-dimensional printed models in temporal bone education. Laryngoscope 2015;125(10): 2353–2357
5 Unger BJ, Kraut J, Rhodes C, Hochman J. Design and Validation of 3D Printed Complex Bone Models with Internal Anatomic Fidelity for Surgical Training and Rehearsal. Stud Health Technol Inform 2014;196:439–445
6 Hochman JB, Rhodes C, Kraut J, Pisa J, Unger B. End user comparison of anatomically matched 3-dimensional printed and virtual haptic temporal bone simulation: a pilot study. Otolaryngol Head Neck Surg 2015;153(02):263–268
7 Hochman JB, Sepehri N, Rampersad V, et al. Mixed reality temporal bone surgical dissector: mechanical design. J Otolaryngol Head Neck Surg 2014;43(01):23
8 Hochman JB, Unger B, Kraut J, Pisa J, Hombach-Klonisch S. Gesture-controlled interactive three dimensional anatomy: a novel teaching tool in head and neck surgery. J Otolaryngol Head Neck Surg 2014;43(01):38
9 Wong V, Unger B, Pisa J, Gousseau M, Westerberg B, Hochman JB. Construct Validation of a Printed Bone Substitute in Otologic Education. Otol Neurotol 2019;40(07):e698–e703
10 Frithioff A, Frendø M, Pedersen DB, Sørensen MS, Wuyts Andersen SA. 3D-Printed Models for Temporal Bone Surgical Training: A Systematic Review. Otolaryngol Head Neck Surg 2021 Nov;165 (05):617–625
11 Lui JT, Hoy MY. Evaluating the Effect of Virtual Reality Temporal Bone Simulation on Mastoidectomy Performance: A Meta-analysis. Otolaryngol Head Neck Surg 2017;156(06):1018–1024
12 Pisa J, Gousseau M, Mowat S, Westerberg B, Unger B, Hochman JB. Simplified Summative Temporal Bone Dissection Scale Demonstrates Equivalence to Existing Measures. Ann Otol Rhinol Laryngol 2018;127(01):51–58
13 Pisa J, Gousseau M, Mowat S, Westerberg B, Unger B, Hochman JB. Simplified Summative Temporal Bone Dissection Scale Demonstrates Equivalence to Existing Measures. Annals of Otology, Rhinology & Laryngology 2018;127(01):51–58
14 Zirkle M, Taplin MA, Anthony R, Dubrowski A. Objective assessment of temporal bone drilling skills. Ann Otol Rhinol Laryngol 2007;116(11):793–798
15 Laeeq K, Bhatti NI, Carey JP, et al. Pilot testing of an assessment tool for competency in mastoidectomy. Laryngoscope2009–12119(12):2402–2410
16 Butler NN, Wiet GJ. Reliability of the Welling scale (WS1) for rating temporal bone dissection performance. Laryngoscope2007–10117(10):1803–1808
17 Mowry SE, Woodson E, Gubbels S, Carfrae M, Hansen MR. A simple assessment tool for evaluation of cadaveric temporal bone dissection. Laryngoscope 2018;128(02):451–455
18 Andersen SAW, Konge L, Mikkelsen PT, Cayé-Thomasen P, Sørensen MS. Mapping the plateau of novices in virtual reality simulation training of mastoidectomy. Laryngoscope 2017;127 (04):907–914
19 Khemani S, Arora A, Singh A, Tolley N, Darzi A. Objective skills assessment and construct validation of a virtual reality temporal bone simulator. Otol Neurotol 2012;33(07):1225–1231

