The Journal of Laryngology & Otology

cambridge.org/jlo

Objective structured assessment of technical skill in temporal bone dissection: validation of a novel tool

M Stavrakas[1], G Menexes[2], S Triaridis[1], P Bamidis[3], J Constantinidis[1] and P D Karkos[1]

[1]ENT Department, AHEPA University Hospital, Aristotle University of Thessaloniki, Thessaloniki, Greece, [2]Department of Field Crops and Ecology, Faculty of Agriculture, Forestry and Natural Environment, School of Agriculture, Aristotle University of Thessaloniki, Thessaloniki, Greece and [3]Laboratory of Medical Physics, Medical School, Aristotle University of Thessaloniki, Thessaloniki, Greece

Main Article

Dr M Stavrakas takes responsibility for the integrity of the content of the paper

Cite this article: Stavrakas M, Menexes G, Triaridis S, Bamidis P, Constantinidis J, Karkos PD. Objective structured assessment of technical skill in temporal bone dissection: validation of a novel tool. *J Laryngol Otol* 2021; **135**:518–528. https://doi.org/10.1017/S0022215121001201

Accepted: 27 October 2020
First published online: 12 May 2021

Key words:
Otolaryngology; Otologic Surgical Procedures; Printing

Author for correspondence:
Dr Marios Stavrakas, ENT Department,
Aristotle University of Thessaloniki,
AHEPA Hospital, Thessaloniki Kiriakidi 1,
Thessaloniki 546 21, Greece
E-mail: mstavrakas@doctors.org.uk

Abstract

Objective. This study developed an assessment tool that was based on the objective structured assessment for technical skills principles, to be used for evaluation of surgical skills in cortical mastoidectomy. The objective structured assessment of technical skill is a well-established tool for evaluation of surgical ability. This study also aimed to identify the best material and printing method to make a three-dimensional printed temporal bone model.

Methods. Twenty-four otolaryngologists in training were asked to perform a cortical mastoidectomy on a three-dimensional printed temporal bone (selective laser sintering resin). They were scored according to the objective structured assessment of technical skill in temporal bone dissection tool developed in this study and an already validated global rating scale.

Results. Two external assessors scored the candidates, and it was concluded that the objective structured assessment of technical skill in temporal bone dissection tool demonstrated some main aspects of validity and reliability that can be used in training and performance evaluation of technical skills in mastoid surgery.

Conclusion. Apart from validating the new tool for temporal bone dissection training, the study showed that evolving three-dimensional printing technologies is of high value in simulation training with several advantages over traditional teaching methods.

Introduction

A hugely important part of surgery is the training of doctors, followed by the assessment of their competence and quality of the training they have received. Improvement of surgical skills should not follow Halsted's model, which claims that learning is achieved by performing the procedure.[1] The principle 'see one, do one, teach one' tends to be abandoned as ineffective.[2] The training methods that simulate real conditions and scenarios have been conscripted in numerous other industries, such as aviation, architecture and the military. Simulation has entered medical education only during the past decade. In order to understand the role of simulation in medical training, it is useful to define the term. Bismuth *et al.* define simulation as 'a person, device or set of conditions which attempts to present [education and] evaluation problems authentically'.[1]

Training in surgery is entirely different from medical training. As a result, some training programmes end up producing less experienced and less competent surgeons owing to the decreased number of training hours. This could be correlated to the fact that every trainee surgeon has a different learning curve. Moreover, some surgeons may finish their training at a lower point on their learning curve.[3] Simulation is an excellent adjunct in training and has been adopted by many surgical specialties, including otolaryngology.[4]

According to a literature review by Musbahi *et al.*, there are 64 otolaryngology simulators available, including virtual reality and bench models, with various levels of validity.[4] The integration of surgical simulation in training is essential as it endorses clinical skill acquisition in an environment of reduced learning opportunities, especially after the introduction of the European Working Time Directive.[5] Moreover, it enhances communication, decision-making processes and situational awareness.[6]

Medical students and specialty trainees are familiar with objective structured clinical examination, which represents a method of assessment of skills in physical examination, communication and professionalism.[2] Although it seems to be a widely accepted method of evaluation, it cannot be applied in surgery, as it does not assess technical skills. The objective structured assessment of technical skill ('OSATS') was developed in Toronto by Martin *et al.*[7] with the purpose of assessing the development of surgical skills.

The objective structured assessment of technical skill in temporal bone dissection ('TempOSATS') is a novel proposed tool. Its principal aim is to assess surgical skills in

https://doi.org/10.1017/S0022215121001201 Published online by Cambridge University Press

Table 1. Comparison of different printing materials and methods

Parameter	Polylactic acid	Polylactic acid + polyvinyl alcohol	Resin-standard	Resin-selective laser sintering
Anatomy	Very good. Well aerated mastoid cells but did not approach anatomy so well	Good	Good	Excellent. Mastoid cells were empty of material
Landmarks	Very good. Not so easily identifiable	Good	Good	Excellent. Surface landmarks easily identifiable
Tactile feedback	Very good. Worse than resin	Good	Excellent	Excellent. Close to realistic
Instrumentation	Good. Material melting due to heat	Good	Excellent	Excellent. Easy instrumentation
Cost	Excellent. Low cost	Excellent	Good	Good. Higher cost/model

temporal bone dissection and more specifically in cortical mastoidectomy, according to the already validated pillars of the objective structured assessment of technical skill tool.

Our study comprised two aims. The first was to assess the best material to make a three-dimensional (3D) temporal bone model, to present the advantages of 3D printing in temporal bone dissection as a means of surgical simulation and to implement these technologies in setting up a skills laboratory using exclusively 3D-printed models. Moreover, we aim to explore some main aspects of the validity and reliability of the proposed objective structured assessment of technical skill in temporal bone dissection tool, which is based on the principles of the objective structured assessment of technical skill, as an assessment tool for basic temporal bone dissection, utilising 3D-printing techniques to establish identical anatomical models.

Materials and methods

Selection of materials and printing modality

After selecting a computed tomography (CT) scan of a well aerated, disease-free temporal bone, we converted the Dicom® data to a stereolithographic ('stl') file, which is appropriate for 3D printing. Only a few improvements were required to limit any artifacts in the final format, such as removal of supporting structures from the mastoid air cells and draining holes, depending on the printing method.

The main question was which 3D-printing technology would approach anatomical accuracy of the real temporal bone, allow quick reproducibility, satisfactory tactile feedback and affordable cost. The materials we tested were polylactic acid, polylactic acid plus polyvinyl alcohol, resin by conventional printing and selective laser sintering.

Afterwards, we assessed all four models by a focus group, consisting of five specialist otolaryngologists with experience in temporal bone surgery. The focus group also agreed on the steps that should be included in the objective structured assessment of technical skill in temporal bone dissection tool. These are the main surgical steps described in the literature and also reflect the experience of the focus group.[8,9] The advantages and disadvantages of these models are summarised in Table 1. Paying attention to the anatomical resemblance and feedback to drilling, we concluded that selective laser sintering resin technology was the best for our purpose. The cost of each model was approximately 25 euros.

The experiments were conducted on a simple bench with a temporal bone holder and drill, which can be easily replaced by a Dremel-type drill (Illinois, USA). Different types of drill heads were available (cutting and diamond), as well as suction, irrigation and otological micro-instruments (for example, needles and crocodile forceps). The task was cortical mastoidectomy. MacEwen's triangle could be easily identified as the spine of Henle and the zygomatic root. Drilling of the selective laser sintering model was smooth, with close to realistic tactile feedback. The mastoid cells were empty of material, and the position of the other landmarks (sigmoid sinus, lateral semi-circular canal and incus buttress) could also be identified. All the surgical steps were previously agreed by the members of the group, executed in an uninterrupted sequence and videotaped so they could be reassessed later (Figure 1).

Selection of sample and simulation process

To determine the minimum required sample, power analysis was conducted following the minimum expected correlation coefficient (Spearman's rank correlation rho) for testing inter- and intra-rater reliability of the two assessors relative to their total scoring of overall achievement. For an anticipated correlation coefficient of rho = 0.60 utilising a sample size of at least $n = 19$ units, a two-tailed t-test for testing the statistical significance of the corresponding correlation coefficient, at significance level $a = 0.05$, showed enough power ($\gamma = 0.80$) to highlight the association as statistically significant. Generally, a value of a correlation coefficient of 0.60 is considered to correspond to a 'large' effect size according to Cohen's conventions.[10] Power analysis was conducted with G*Power (version 3.1.2) statistical power analysis software (software detailed in Faul *et al.*[11] and Faul *et al.*[12]).

A flowchart of the methodology and the experimental part is presented in Figure 1. Two of the authors acted as external assessors, who initially delivered a brief tutorial to the candidate (slides disseminated via e-mail), focusing on the objectives and surgical steps that were expected to be performed. Following this, specialty trainees of various levels from rotations in Northern Greece were asked to perform a cortical mastoidectomy in the pilot skills labarotary using the selective laser sintering resin printed temporal bone models. They were invited via personal e-mail invitation, and their participation was registered on a first come, first-served basis. They all had the same equipment available to complete the task, and they were videotaped (Figures 2–4).

Assessment and scoring

The videos were given a number from 1 to 24. Then they were scored according to objective structured assessment of technical skill in temporal bone dissection by the two external assessors at two different times: after the completion of the

https://doi.org/10.1017/S0022215121001201 Published online by Cambridge University Press

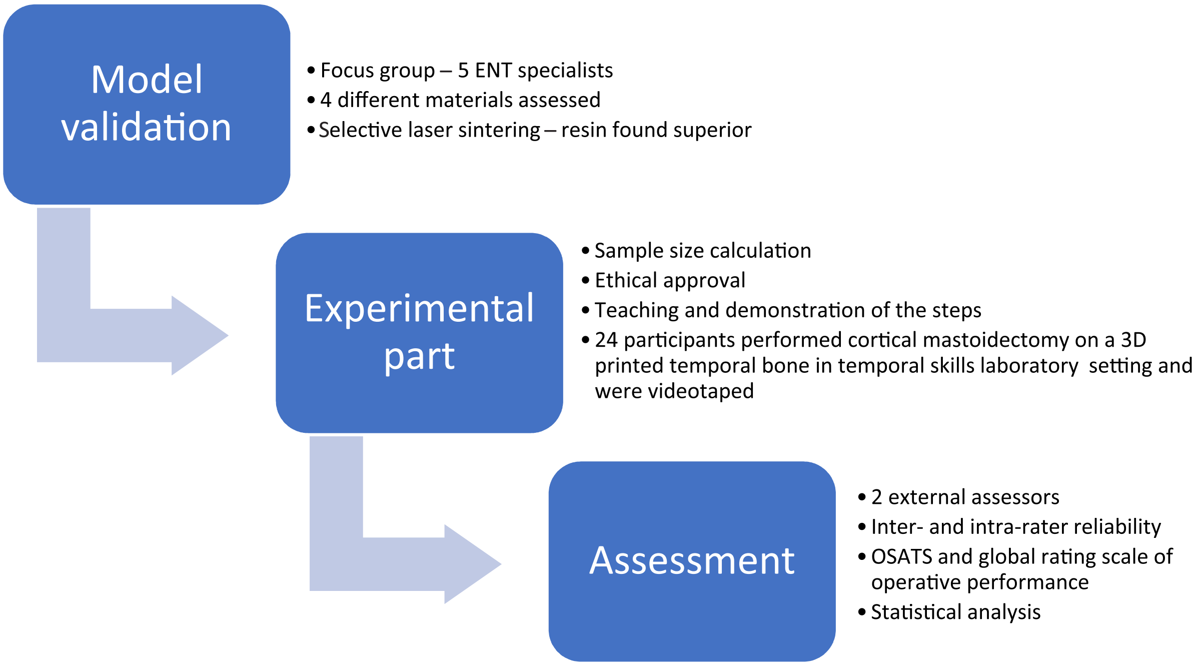

Model validation
• Focus group – 5 ENT specialists
• 4 different materials assessed
• Selective laser sintering – resin found superior
Experimental part
• Sample size calculation
• Ethical approval
• Teaching and demonstration of the steps
• 24 participants performed cortical mastoidectomy on a 3D printed temporal bone in temporal skills laboratory setting and were videotaped
Assessment
• 2 external assessors
• Inter- and intra-rater reliability
• OSATS and global rating scale of operative performance
• Statistical analysis

Fig. 1. Flowchart of methodology. 3D = three dimensional; OSATS = objective structured assessment for technical skills

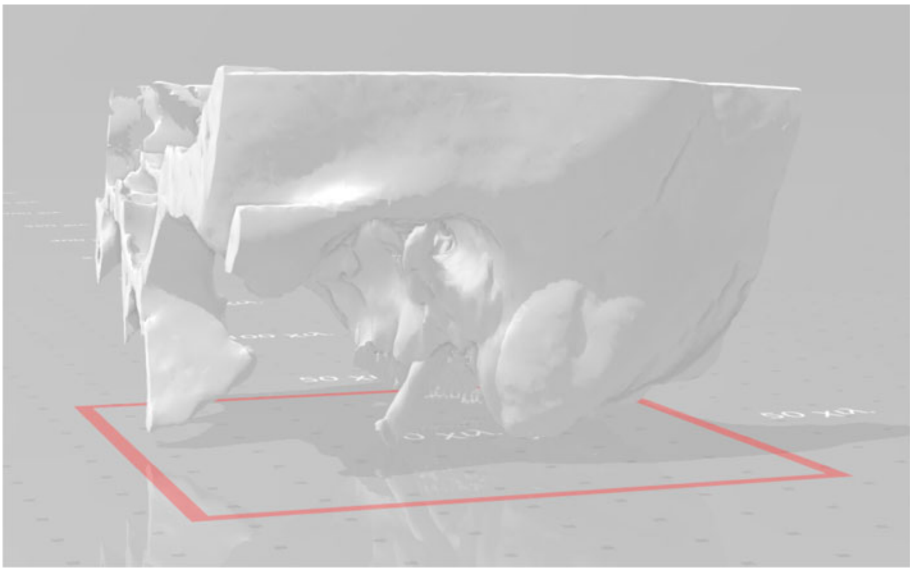


Fig. 2. Temporal bone three-dimensional model.

experimental part and one month later. Before scoring, a meeting took place for calibration purposes, and the two assessors agreed on the scoring methodology. According to the literature, similar projects involved 2–3 assessors, directly evaluating the candidates, especially for the first time.[7,13,14] Additionally, the videos were reviewed again after one month. This is in line with the relevant literature, where intra-rater variability was assessed by reviewing video recordings after some days up to six weeks.[14–17]

The importance of video recording has been highlighted in several studies.[18,19] The assessors also scored the candidates according to an already validated global rating scale,[7,20] which was utilised as a control tool for testing the criterion validity of the proposed objective structured assessment of technical skill in temporal bone dissection. As shown in Figure 5, objective structured assessment of technical skill in temporal bone dissection consists of seven questions (scored as yes/no) and one question of overall achievement, scored from 0 to 5. The global rating scale has seven questions, scored from one to five (Figure 6).

The study was approved by the Committee of Bioethics of the Aristotle University Medical School, Thessaloniki, Greece. All participants gave written consent before participating in the experimental part, and the consent forms were also approved by the Committee of Bioethics.

Statistical analysis

Data were summarised by calculating descriptive statistical indices such as absolute and relative frequencies (percentages), measures of central tendency (means and medians) and variability (standard deviations), correlation-association indices (Spearman's rho for correlating quantitative variables, and

https://doi.org/10.1017/S0022215121001201 Published online by Cambridge University Press

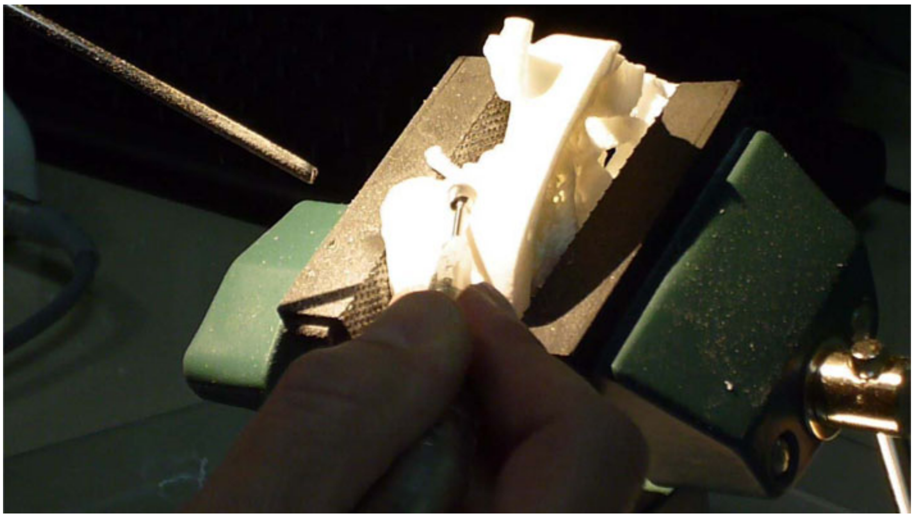


Fig. 3. Temporal bone skills station.

gamma or Cramer's V for assessing the degree of the association between categorical variables).

The process of testing some aspects of the reliability and the validity of the proposed objective structured assessment of technical skill in temporal bone dissection assessment tool was based on the following methodological scheme: (1) the internal consistency of the objective structured assessment of technical skill in temporal bone dissection tool was tested and evaluated by estimating and assessing the value of the Kuder–Richardson formula 20 reliability coefficient.[21,22] The Kuder–Richardson formula 20 coefficient is analogous to Cronbach's a reliability coefficient, but it is appropriate for binary items. (2) For both tools, the average discrimination index was calculated. The discrimination index was used for testing the homogeneity of the two tools.[21] This index is related mainly to the construct validity of a scale consisting of several items. These first two analyses were performed for each examiner within each evaluation time (time 1 and time 2). (3) The criterion validity of the objective structured assessment of technical skill in temporal bone dissection assessment tool was tested and evaluated by correlating, at each evaluation time (time 1 and time 2), the examiners' scores on the overall assessment item of the objective structured assessment of technical skill in temporal bone dissection tool with the average score of the global rating scale of operative performance tool. (4) The 'inter-rater' and 'intra-rater' reliability were tested with Spearman's rho and Wilcoxon tests.

In all statistical tests, the observed significance level (p-value) was computed with the Monte-Carlo simulation method utilising 10 000 random samples.[23,24] All the statistical analyses were performed with SPSS® (version 24.0) statistical software enhanced with the module 'exact tests' (for the implementation of the Monte-Carlo simulation). The significance level in all hypothesis testing procedures was predetermined at $a = 0.05$ ($p \leq 0.05$).

Results

According to data presented in Table 2, the vast majority of the scores of the two examiners using the two tools, for both time periods, showed satisfactory reliability indices (Kuder–Richardson formula 20 or Cronbach's a reliability coefficients more than or equal to 0.60) and homogeneity (average discrimination index more than 0.30).

Based on data presented in Tables 3 and 4, for each examiner, there was a very strong (almost absolute) positive and statistically significant correlation between examiner scores at time 1 and time 2 for the overall assessment of objective structured assessment of technical skill in temporal bone dissection tool (for examiner one: rho = 0.942, $p < 0.001$; for examiner 2: rho = 0.908, $p < 0.001$). However, for examiner one there was a statistically significant difference ($p = 0.002$) between the two assessments (time 1 *vs* time 2). The mean value of the overall evaluation at time 1 was estimated to be 3.8 and at time 2 was estimated to be 3.3; that is, significantly lower than time 1 (mean difference was equal to 0.5 in a 6-point scale). For examiner two, no statistically significant difference ($p = 0.748$) between the two assessments was highlighted, according to the results of the Wilcoxon test. It must be noted that in all comparisons, the median values were all equal to 4.0.

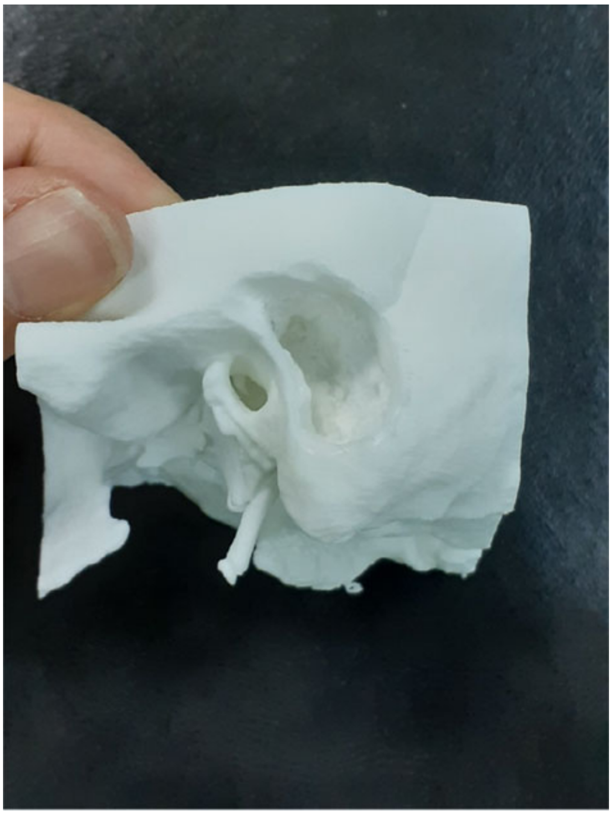
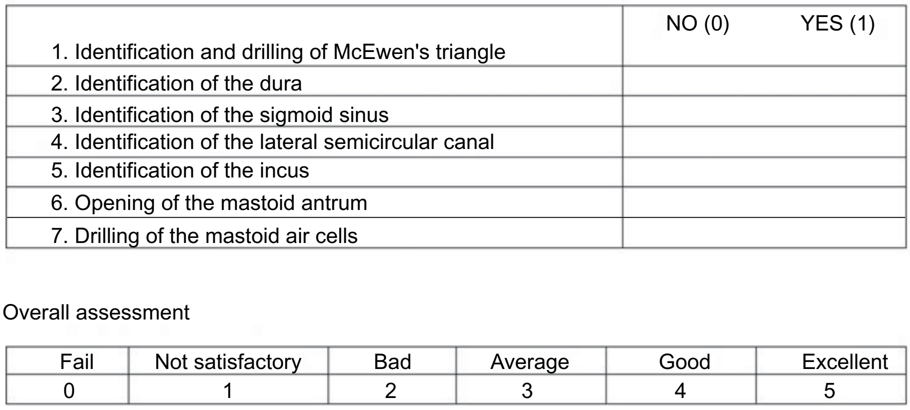


Fig. 4. Three-dimensional printed model after cortical mastoidectomy.

https://doi.org/10.1017/S0022215121001201 Published online by Cambridge University Press



	NO (0)	YES (1)
1. Identification and drilling of McEwen's triangle		
2. Identification of the dura		
3. Identification of the sigmoid sinus		
4. Identification of the lateral semicircular canal		
5. Identification of the incus		
6. Opening of the mastoid antrum		
7. Drilling of the mastoid air cells		

Overall assessment

Fail	Not satisfactory	Bad	Average	Good	Excellent
0	1	2	3	4	5

Fig. 5. The objective structured assessment of technical skill in temporal bone dissection tool for assessment of cortical mastoidectomy.

According to the data presented in Tables 3 and 4, at each time point there was a very strong (almost absolute at time 2) positive and statistically significant correlation between the scores of the two examiners for the overall assessment of the objective structured assessment of technical skill in temporal bone dissection tool (at time 1: rho = 0.837, $p < 0.001$; at time 2: rho = 0.999, $p < 0.001$). However, at time 1, there was a statistically significant difference ($p = 0.035$) between the two examiners. The mean value of the overall assessment for examiner one was estimated to 3.8, and for examiner two, it was equal to 3.4. That is, significantly lower than examiner one (mean difference was equal to 0.4 on a 6-point scale). At time 2, no statistically significant difference ($p = 1.000$) between the two examiners was found, according to the results of the Wilcoxon test.

Based on data presented in Tables 5 and 6, for both examiners at times 1 and 2, there was a very strong, positive and statistically significant correlation ($p < 0.001$) between their overall assessment scores derived from the two tools.

Landis and Koch (1977) remark that kappa values around 0.20 express a weak degree of agreement, values around 0.40 indicate a satisfactory degree of agreement, values around 0.60 express a moderate degree of agreement, values around 0.80 indicate a significant degree of agreement and, finally, kappa values over 0.80 express an almost perfect degree of agreement.[25] Based on the data presented in Table 7, the vast majority of Cohen's kappa measures of agreement were greater than 0.80 and statistically significant (maximum $p = 0.042$, <0.05). The simple overall agreement percentages between any two assessments' scores were greater than 95 per cent (ranged from 96 to 100 per cent). Regarding the degree of the association between any two assessments' scores, the corresponding association indices were very high (both Cramer's V > 0.80 and gamma > 0.80, range, 0.836 to 1) and statistically significant ($p < 0.001$). Consequently, testing the items of the objective structured assessment of technical skill in temporal bone dissection tool, the two examiners showed very strong agreement between their intra- and inter-reliability assessments.

Tables 8 and 9 present the results of the intra- and inter-rater reliability testing for the average summated score of the global rating scale of operative performance tool. In all testing procedures, there was a very strong (almost absolute) positive and statistically significant correlation between any two assessments, in all cases $p < 0.001$ (Tables 8 and 9).

Discussion

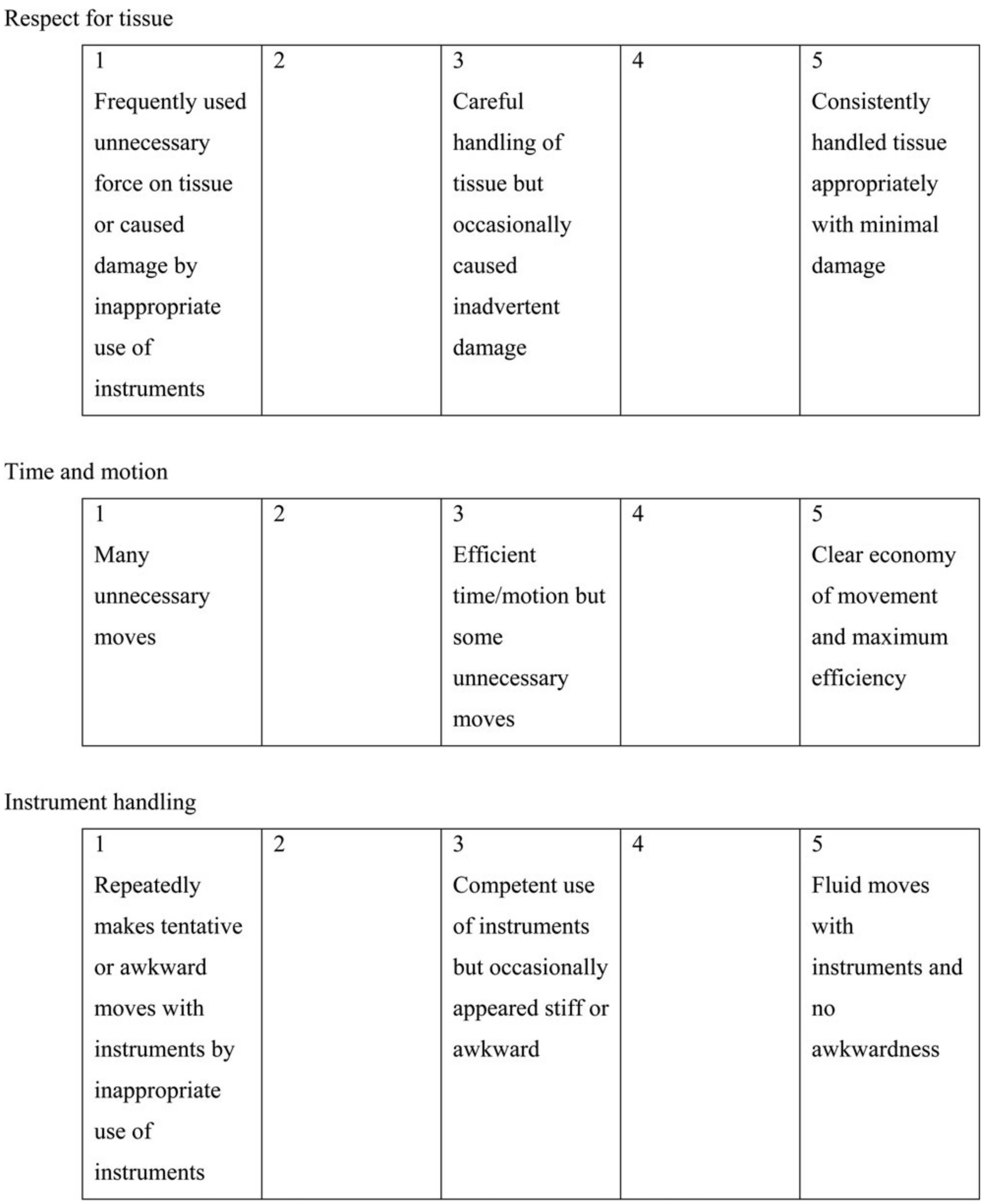
Surgical training was previously mainly confined to the practice and development of surgical skills in the operating theatre. According to Reznick *et al.*, the operating theatre has many limitations when it comes to training and assessment. First of all, it is difficult to standardise any operation in similar training patterns. Secondly, it is almost impossible to standardise the degree to which a trainee is performing elements of an operation. In addition, surgical time is far more expensive compared with any other training method.[20]

Aiming to overcome the above limitations, efforts have been made to develop effective teaching methods. Animal models are carefully selected to simulate human anatomy, and the animal must be anaesthetised before the operation. Obviously, ethical issues are involved, and animal models do not offer a wide range of alternatives to real patients. The use of bench models simulates human anatomy well and are used for ordinary surgical tasks. Comparing bench model based training to previous methods, it has a lower cost, is portable, readily available and allows the reproducibility of various tasks.[7]

Objective structured assessment of technical skill gives the candidate a score that ranges from 8 to 40, with 24 representing a competent performance. Pandey *et al.*[3] described the value of objective structured assessment of technical skill. Despite the small number of participants (15 surgical trainees), this study showed that the participants had significant improvement in all aspects of their generic skill but mainly improved in the flow of the procedure, their overall performance and their procedure-specific skills. In the same study, although significant improvement was observed, some participants did not improve. They were mainly older surgeons who proved to be less able to learn in this type of setting because they had accumulated other methods of performing the examined procedures. Another reason may be that they have learned other types of the same procedure that are different from those demonstrated to them.[3]

The Vascular Department of Imperial College London, which is based at St Mary's Hospital, adopted objective structured assessment of technical skill in their surgeons' training. They took objective structured assessment of technical skill a step beyond its original idea: evaluating surgical competence in a specific procedure and not only basic surgical tasks. The new tool that Imperial introduced was called Imperial College Evaluation of Procedure Specific Skill. This involves a rating scale with five standard points to assess the content of a procedure.[1]

There is no doubt as to the value of surgical skills assessment. The most beneficial impact is the considerable improvement in patient safety because the trainee surgeon does not practice a specific procedure on a patient for the first time.



Respect for tissue

1	2	3	4	5
Frequently used unnecessary force on tissue or caused damage by inappropriate use of instruments		Careful handling of tissue but occasionally caused inadvertent damage		Consistently handled tissue appropriately with minimal damage

Time and motion

1	2	3	4	5
Many unnecessary moves		Efficient time/motion but some unnecessary moves		Clear economy of movement and maximum efficiency

Instrument handling

1	2	3	4	5
Repeatedly makes tentative or awkward moves with instruments by inappropriate use of instruments		Competent use of instruments but occasionally appeared stiff or awkward		Fluid moves with instruments and no awkwardness

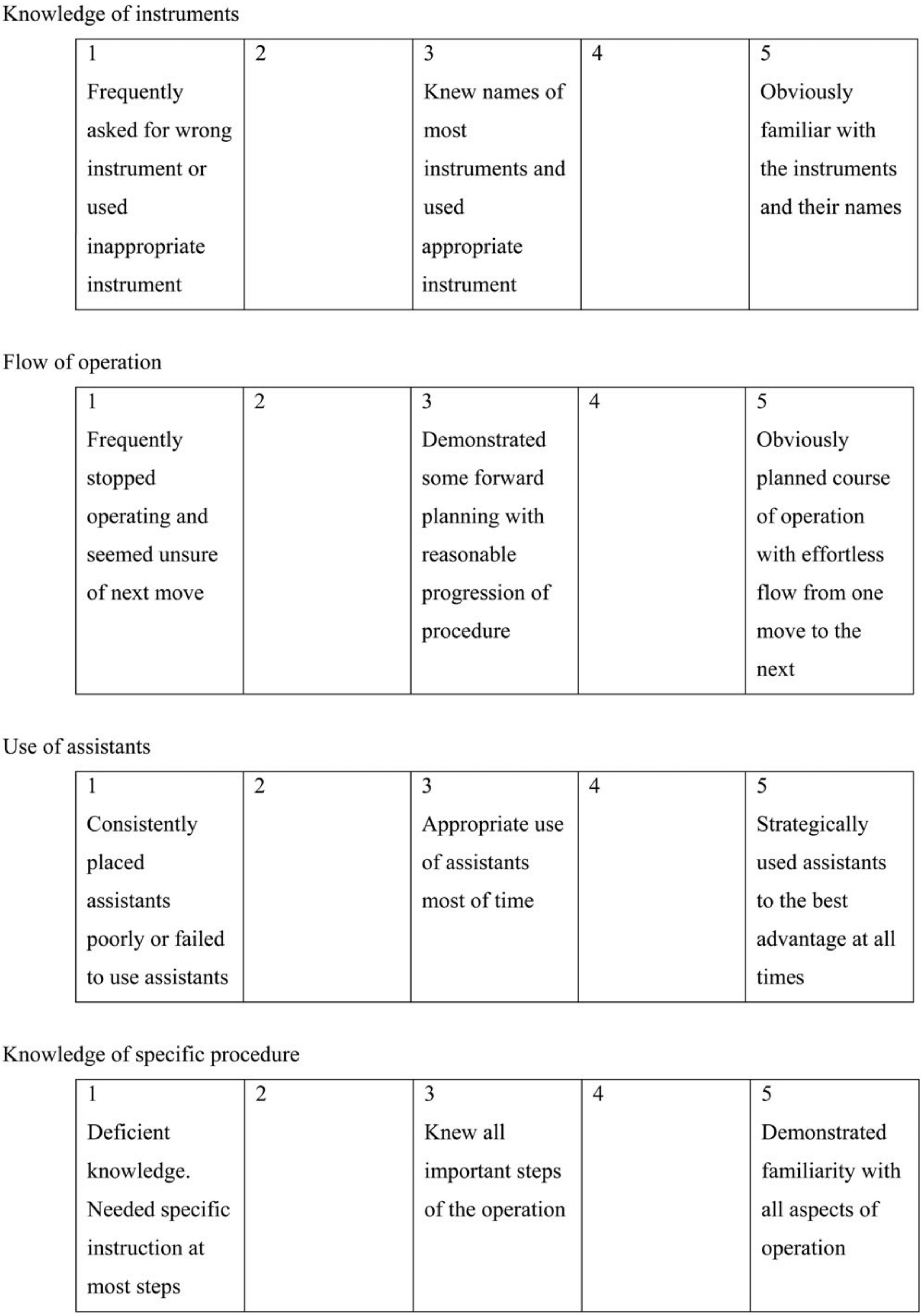
Fig. 6. Global rating scale.

Continued.

In addition, the 'learning curve' of making mistakes takes place in the laboratory and not on a patient. In that way, the trainee can perform the same procedure many times until improvement is reached. As a result, operating time decreases, efficiency increases and medical errors decrease.[2] This agrees with our philosophy of applying the objective structured assessment of technical skill principles to the whole surgical procedure and not only for limited skills. Moreover, our experiments demonstrated the need for adequate calibration between the assessors, some discrepancies in scoring that may have to do with the different levels of experience of the assessors and the value of video recordings, which allow more careful evaluation of the various surgical steps.

A possible problem in applying objective structured assessment of technical skills in every training hospital is the relatively high cost. When the method first became known, only a few major teaching centres had the resources to organise courses and evaluations, and this could only occur a few

https://doi.org/10.1017/S0022215121001201 Published online by Cambridge University Press



Knowledge of instruments

1	2	3	4	5
Frequently asked for wrong instrument or used inappropriate instrument		Knew names of most instruments and used appropriate instrument		Obviously familiar with the instruments and their names

Flow of operation

1	2	3	4	5
Frequently stopped operating and seemed unsure of next move		Demonstrated some forward planning with reasonable progression of procedure		Obviously planned course of operation with effortless flow from one move to the next

Use of assistants

1	2	3	4	5
Consistently placed assistants poorly or failed to use assistants		Appropriate use of assistants most of time		Strategically used assistants to the best advantage at all times

Knowledge of specific procedure

1	2	3	4	5
Deficient knowledge. Needed specific instruction at most steps		Knew all important steps of the operation		Demonstrated familiarity with all aspects of operation

Fig. 6. Continued.

https://doi.org/10.1017/S0022215121001201 Published online by Cambridge University Press

Table 2. Reliability results of the two tools used by the two assessors at two time points

	Time 1				Time 2			
	Examiner 1		Examiner 2		Examiner 1		Examiner 2	
Assessment scale	KR20/ Cronbach's *α*	DI	KR20/ Cronbach's *α*	DI	KR20/ Cronbach's *α*	DI	KR20/ Cronbach's *α*	DI
TempOSATS (7 yes/no items)	0.513	0.259	0.730	0.471	0.637	0.476	0.718	0.399
Global rating scale of operative performance (7 5-point ordinal scale items)	0.960	0.878	0.974	0.923	0.967	0.911	0.977	0.948

For the objective structured assessment of technical skill in temporal bone dissection (TempOSATS) tool items, Cronbach's *α* reliability coefficient is equivalent to Kuder–Richardson formula 20 (KR20) reliability coefficient, and discrimination index (DI) is the average discrimination index

Table 3. TempOSATS overall assessment intra-rater reliability and comparison of means

	Time 1 score		Time 2 score			
Examiner	Mean	Median	Mean	Median	Correlation (rho (*p*-value))	Wilcoxon *p*-value
Examiner 1	3.8	4.0	3.3	4.0	0.942 (<0.001)	0.002
Examiner 2	3.4	4.0	3.3	4.0	0.908 (<0.001)	0.746

Table shows intra-rater reliability (Spearman's rho rank correlation coefficient) and comparison of means, for each examiner, between time 1 and time 2 and between the two examiners at each time point, for the objective structured assessment of technical skill in temporal bone dissection (TempOSATS) score for overall assessment

Table 4. TempOSATS overall assessment of intra-rater reliability and comparison of means

Time point	Examiner 1 score (mean)	Examiner 2 score (mean)	Correlation (rho (*p*-value))	Wilcoxon *p*-value
Time 1	3.8	3.4	0.837 (<0.001)	0.035
Time 2	3.3	3.3	0.999 (<0.001)	1.000

Table shows intra-rater reliability (Spearman's rho rank correlation coefficient) and comparison of means, for each examiner, between time 1 and time 2 and between the two examiners at each time point, for the objective structured assessment of technical skill in temporal bone dissection (TempOSATS) score for overall assessment

Table 5. Correlation between the overall assessment scores of the two tools at time 1

Parameter	TempOSATS overall assessment for E1T1 (rho (*p*-value))	TempOSATS overall assessment for E2T1 (rho (*p*-value))
Average summated score of global rating scale of operative performance (E1T1)	0.849 (<0.001)	
Average summated score of global rating scale of operative performance (E2T1)		0.928 (<0.001)

Table shows correlation (Spearman's rho rank correlation coefficient) between the overall assessment scores of the two tools, objective structured assessment of technical skill in temporal bone dissection (TempOSATS) and global rating scale of operative performance, reported by the two examiners at time 1. E1T1 = examiner 1 at time 1; E2T1 = examiner 2 at time 1

times a year. Cost for models, facilities and especially trainers are obstacles to its wider spread.[3] In our study, we managed to reproduce a number of identical models of temporal bones at a low cost, and the printing time was a few hours for each.

Three-dimensional printing is a technology that has been known since the 1980s, but its involvement in the medical field has increased significantly over the last two decades, with numerous examples in training, patient education and bioengineering. Three-dimensional printing equipment has improved, is less expensive and the expertise is more widespread, and therefore it has become available in many parts of the world for medical use in several fields.[26,27]

There are numerous studies available in the literature, exploring the potential use of 3D-printing technologies in ENT head and neck surgery. They vary from pre-operative planning and patient education to more advanced training applications for residents and undergraduate medical students. Additionally, there have been descriptions of applications associated with tissue engineering and prosthetics, which are extremely promising for medical innovations in the near future.

According to Canzi *et al.*, there are 23 studies in the literature focusing on otological applications in training, mainly to do with temporal bone surgery simulation.[28] In 2015, a temporal bone model based on CT scan data of two selected patients with well pneumatised and disease-free mastoids was developed. The final evaluation of the models showed satisfactory reproducibility of most structures and anatomical landmarks but also raised two significant issues: the accuracy of the ossicular chain (mainly the stapes) and also the retained resin within the mastoid air cells. The latter issue impacts the drilling experience and can be overcome by adding a small drain hole in the region of the sigmoid sinus. The authors concluded that the model produced is useful for training, without depleting a limited supply of cadavers and by using conventional (non-surgical) tools, such as a Dremel drill.[29] On the

Table 6. Correlation between the overall assessments scores of the two tools at time 2

Parameter	TempOSATS overall assessment (E1T2) (rho (*p*-value))	TempOSATS overall assessment (E2T2) (rho (*p*-value))
Average summated score of global rating scale of operative performance (E1T2)	0.964 (<0.001)	
Average summated score of global rating scale of operative performance (E2T2)		0.971 (<0.001)

Table shows correlation (Spearman's rho rank correlation coefficient) between the overall assessment scores of the two tools, objective structured assessment of technical skill in temporal bone dissection (TempOSATS) and global rating scale of operative performance, reported by the two examiners at time 2. E1T2 = examiner 1 at time 2; E2T2 = examiner 2 at time 2

Table 7. Degree of agreement or correlation between scores and overall performance of TempOSATS tool

Parameter	Time 1 E1E2	Time 2 E1E2	Examiner 1 T1T2	Examiner 2 T1T2
Identification and drilling of McEwen's triangle (% (*n*))	100 (24/24)*	96 (23/24)*	100 (24/24)*	96 (23/24)*
Identification of the dura (kappa (*p*-value))	0.895 (<0.001)	0.903 (<0.001)	0.895 (<0.001)	0.903 (<0.001)
Identification of the sigmoid sinus (kappa (*p*-value))	0.780 (<0.001)	0.710 (0.002)	0.780 (<0.001)	0.903 (<0.001)
Identification of the lateral semicircular canal (kappa (*p*-value))	0.913 (<0.001)	0.829 (<0.001)	0.829 (<0.001)	0.913 (<0.001)
Identification of the incus (kappa (*p*-value))	0.834 (<0.001)	0.913 (0.001)	0.753 (<0.001)	1.000 (<0.001)
Opening of the mastoid antrum ((% (*n*) or (kappa (*p*-value))	96 (23/24)*	96 (23/24)*	100 (24/24)*	Kappa = 1.000 (*p* = 0.042)
Opening of the mastoid air cells ((% (*n*) or (kappa (*p*-value))	96 (23/24)*	96 (23/24)*	Kappa = 1.000 (*p* = 0.041)	100 (24/24)*
Overall performance				
– Cramer's V (*p*-value)	0.836 (<0.001)†	1.000 (<0.001)†	0.888 (<0.001)†	0.903 (<0.001)†
– Gamma (*p*-value)	0.896 (<0.001)†	1.000 (<0.001)†	1.000 (<0.001)†	0.915 (<0.001)†

Table shows degree of agreement (Cohen's kappa measure) or correlation (Cramer's V and gamma association indices) between the two examiners' scores within each attempt and between the two attempts (time 1 and time 2) for the 7 items and the overall performance of the objective structured assessment of technical skill in temporal bone dissection (TempOSATS) assessment tool. *In those cases where it was not possible to compute the Cohen's kappa measure of agreement, the simple overall agreement percentage between any two assessments is reported instead; †in those cases where it was not possible to compute the Cohen's kappa measure of agreement, the Cramer's V and gamma association indices between any two assessments are reported instead. E1 = examiner 1; E2 = examiner 2; T1 = time 1; T2 = time 2

Table 8. Intra-rater reliability and comparison of means for the average summated score of the GRSOP

Average summated score of GRSOP tool for each examiner	Time 1		Time 2			
	Mean	Median	Mean	Median	Correlation (rho (*p*-value))	Wilcoxon *p*-value
Examiner 1	3.8	3.6	3.5	3.6	0.942 (<0.001)	0.008
Examiner 2	3.5	3.6	3.4	3.6	0.984 (<0.001)	0.298

Table shows intra-rater reliability (Spearman's rho rank correlation coefficient) and comparison of means, for each examiner, between time 1 and time 2, and between the two examiners at each time point, for the average summated score of the global rating scale of operative performance (GRSOP).

Table 9. Intra-rater reliability and comparison of means for the average summated score of the GRSOP

Time	Examiner 1		Examiner 2			
	Mean	Median	Mean	Median	Correlation (rho (*p*-value))	Wilcoxon *p*-value
Time 1	3.8	3.6	3.5	3.6	0.946 (<0.001)	<0.001
Time 2	3.5	3.6	3.4	3.6	0.994 (<0.001)	0.004

Table shows intra-rater reliability (Spearman's rho rank correlation coefficient) and comparison of means, for each examiner, between time 1 and time 2, and between the two examiners at each time point, for the average summated score of the global rating scale of operative performance (GRSOP).

other hand, it is still difficult to approach the 'natural' structure of the cadaveric specimen, mainly because of the 'stair-stepping' artifact and the lack of anatomical elements such as the dura, nerves, blood vessels, tympanic membrane, and oval and round windows.[30] We have overcome the obstacles of stair-stepping and retained resin by comparing different materials and printing techniques and choosing selective laser sintering printing. This method allows more accurate printing without retained material and better external and internal contours.

- Objective structured assessment of technical skill is a widely accepted tool for assessing surgical skills
- Only a few of its applications in otolaryngology have been explored so far
- There are numerous studies in the literature exploring the potential use of three-dimensional printing
- Three-dimensional printing is a novel but reliable approach to surgical simulation
- This study explored the validity and reliability of a newly proposed assessment model for surgical training
- The objective structured assessment of technical skill in temporal bone dissection is a tool that can be useful in training assessment

Other groups also confirmed the similarity to the cadaveric specimens and the positive feedback from the trainees.[31–34] More specifically, Hochman *et al.* showed that tactile feedback is satisfactory by analysing subjective and objective methods. The improvement of materials has provided a better simulation of bone consistency, resulting in a more realistic experience.[35] A useful adjunct in training is the coupling with electronic simulators, which offers the possibility of real-time alert in case of vital structural injury. An example is the ElePhant model (Electronic Phantom), where the facial nerve is replaced with a conductive alloy or fibre-optic material, allowing immediate feedback.[36] Anecdotal feedback from the participants confirmed the satisfactory tactile feedback, which is associated with the different thickness of the structures (mastoid air cells and bony labyrinth).

Our group has studied the different materials and printing techniques and the application in relatively larger scales has shown that such methods can be used to run skills labs based on 3D-printed models.

Conclusion

Three-dimensional printing is a novel but equally reliable approach to surgical simulation, and reproduction of anatomical models can be of great value in training and personalised patient care. Additionally, objective structured assessment of technical skill in temporal bone dissection is a tool that can be extremely useful in the assessment of training and monitoring of a surgeon's learning curve. More studies are necessary to expand its applications in more complex operations, where cortical mastoidectomy represents the initial stage of surgery.

Competing interests. None declared

References

1 Bismuth J, Donovan MA, O'Malley MK, El Sayed HF, Naoum JJ, Peden EK *et al.* Incorporating simulation in vascular surgery education. *J Vasc Surg* 2010;**52**:1072–80

2 Satava RM. The revolution in medical education-the role of simulation. *J Grad Med Educ* 2009;**1**:172–5

3 Pandey VA, Black SA, Lazaris AM, Allenberg JR, Eckstein HH, Hagmüller GW *et al.* Do workshops improve the technical skill of vascular surgical trainees? *Eur J Vasc Endovasc Surg* 2005;**30**:441–7

4 Musbahi O, Aydin A, Al Omran Y, Skilbeck CJ, Ahmed K. Current status of simulation in otolaryngology: a systematic review. *J Surg Educ* 2017;**74**:203–15

5 Fitzgerald JEF, Caesar BC. The European working time directive: a practical review for surgical trainees. *Int J Surg* 2012;**10**:399–403

6 Yule S, Parker SH, Wilkinson J, McKinley A, MacDonald J, Neill A *et al.* Coaching non-technical skills improves surgical residents' performance in a simulated operating room. *J Surg Educ* 2015;**72**:1124–30

7 Martin JA, Regehr G, Reznick R, Macrae H, Murnaghan J, Hutchison C *et al.* Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997;**84**:273–8

8 Arnoldner C, Lin VYW, Chen JM. Cortical mastoidectomy. In: *Manual of Otologic Surgery*. Vienna: Springer, 2015;5–13

9 Francis HW, Masood H, Laeeq K, Bhatti NI. Defining milestones toward competency in mastoidectomy using a skills assessment paradigm. *Laryngoscope* 2010;**120**:1417–21

10 Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. New Jersey: Lawrence Erlbaum Associates, 1988;13

11 Faul F, Erdfelder E, Lang A-G, Buchner A. G* Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods* 2007;**39**:175–91

12 Faul F, Erdfelder E, Buchner A, Lang A-G. Statistical power analyses using G* Power 3.1: tests for correlation and regression analyses. *Behav Res Methods* 2009;**41**:1149–60

13 Hopmans CJ, den Hoed PT, van der Laan L, van der Harst E, van der Elst M, Mannaerts GHH *et al.* Assessment of surgery residents' operative skills in the operating theater using a modified objective structured assessment of technical skills (OSATS): a prospective multicenter study. *Surgery* 2014;**156**:1078–88

14 Chang OH, King LP, Modest AM, Hur H-C. Developing an objective structured assessment of technical skills for laparoscopic suturing and intracorporeal knot tying. *J Surg Educ* 2016;**73**:258–63

15 Siddiqui NY, Stepp KJ, Lasch SJ, Mangel JM, Wu JM. Objective structured assessment of technical skills for repair of fourth-degree perineal lacerations. *Am J Obstet Gynecol* 2008;**199**:676

16 Siddiqui NY, Galloway ML, Geller EJ, Green IC, Hur H-C, Langston K *et al.* Validity and reliability of the robotic objective structured assessment of technical skills. *Obstet Gynecol* 2014;**123**:1193

17 Schlager A, Ahlqvist K, Rasmussen-Barr E, Bjelland EK, Pingel R, Olsson C *et al.* Inter-and intra-rater reliability for measurement of range of motion in joints included in three hypermobility assessment methods. *BMC Musculoskelet Disord* 2018;**19**:376

18 Jokinen E, Mikkola TS, Härkki P. Simulator training and residents' first laparoscopic hysterectomy: a randomized controlled trial. *Surg Endosc* 2020;**34**:4874–82

19 Rezniczek GA, Severin S, Hilal Z, Dogan A, Krentel H, Buerkle B *et al.* Surgical performance of large loop excision of the transformation zone in a training model: a prospective cohort study. *Medicine* (*Baltimore*) 2017;**96**:7026

20 Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg* 1997;**173**:226–30

21 Nunnally JC. *Psychometric Theory*, 3rd edn. New York: Tata McGraw-Hill Education, 1994

22 Spector PE. *Summated Rating Scale Construction: An Introduction*. Newbury Park, CA: Sage, 1992

23 Mehta CR. *SPSS Exact Tests 7.0 for Windows*. Chicago: SPSS Inc, 1996

24 Mehta CR, Patel NR. Exact permutational inference for categorical and nonparametric data. *Stat Strateg Small Sample Res*. 1999;1–29

25 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74

26 Crafts TD, Ellsperman SE, Wannemuehler TJ, Bellicchi TD, Shipchandler TZ, Mantravadi A V. Three-dimensional printing and its applications in otorhinolaryngology--head and neck surgery. *Otolaryngol Neck Surg* 2017;**156**:999–1010

27 Gross BC, Erkal JL, Lockwood SY, Chen C, Spence DM. Evaluation of 3D printing and its potential impact on biotechnology and the chemical sciences. *Anal Chem* 2014;**86**:3240–53

28 Canzi P, Magnetto M, Marconi S, Morbini P, Mauramati S, Aprile F *et al.* New frontiers and emerging applications of 3D printing in ENT surgery: a systematic review of the literature. *Acta Otorhinolaryngol Ital* 2018;**38**:286–303

29 Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC *et al.* User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* 2006;**31**:1116–28

30 Cohen J, Reyes SA. Creation of a 3D printed temporal bone model from clinical CT data. *Am J Otolaryngol* 2015;**36**:619–24

31 Da Cruz MJ, Francis HW. Face and content validation of a novel three-dimensional printed temporal bone for surgical skills development. *J Laryngol Otol* 2015;**129**:S23–9

32 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**:2353–7

https://doi.org/10.1017/S0022215121001201 Published online by Cambridge University Press

33 Mowry SE, Jammal H, Myer IV C, Solares CA, Weinberger P. A novel temporal bone simulation model using 3D printing techniques. *Otol Neurotol* 2015;**36**:1562–5

34 Rose AS, Webster CE, Harrysson OLA, Formeister EJ, Rawal RB, Iseli CE. Preoperative simulation of pediatric mastoid surgery with 3D-printed temporal bone models. *Int J Pediatr Otorhinolaryngol* 2015;**79**:740–4

35 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 Neck Surg* 2014;**150**:448–54

36 Grunert R, Strauss G, Moeckel H, Hofer M, Poessneck A, Fickweiler U *et al.* ElePhant--an anatomical electronic phantom as simulation--system for otologic surgery. *Conf Proc IEEE Eng Med Biol Soc* 2006;4408–11

https://doi.org/10.1017/S0022215121001201 Published online by Cambridge University Press